

**COMBINING DATA ENVELOPMENT ANALYSIS WITH MACHINE LEARNING
ALGORITHMS FOR PREDICTIONS**

KNUST

By

APPIAHENE, PETER- MPhil.

A Thesis submitted to the Department of Computer Science, Kwame Nkrumah
University of Science and Technology, Kumasi in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

Department of Computer Science

College of Science

September, 2020

**KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY,
KUMASI**

DEPARTMENT OF COMPUTER SCIENCE



**COMBINING DATA ENVELOPMENT ANALYSIS WITH MACHINE LEARNING
ALGORITHM FOR PREDICTIONS**

Author

APPIAHENE PETER
(MPhil.)

Supervisors

Dr. Yaw Marfo Missah

Professor Michael Asante

Doctoral Thesis Submitted to the College of Science in Partial Fulfillment of the
Requirements for the Award of a Degree of

Doctor of Philosophy in Computer Science

© Appiahene, Peter, 2020

DECLARATION

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma at Kwame Nkrumah University of Science and Technology, Kumasi or any other educational institution, except where due acknowledgment is made in the thesis.

APPIAHENE, PETER (20477593)

Name of Student Name and ID

.....
Signature

.....
Date

Certified by:

DR. YAW MARFO MISSAH

Name of First Supervisor

.....
Signature

.....
Date

Certified by:

PROF. MICHAEL ASANTE

Name of Second Supervisor

.....
Signature

Date

Certified by:

PROF. JAMES BEN HAYFRON-ACQUAH

Name of Head of Department

.....
Signature

Date

.....

.....

DEDICATION

This work is dedicated with the deepest respect and honour to my late Uncle, Hon. Charles Yeboah (Former District Chief Executive Atwima), my parents Miss. Sussana Yeboah and Mr.

Alex Opoku Mensah all of blessed memory; my son, Ache Opoku Appiahene, my daughter, Mercedes Yeboah Appiahene and my lovely and understanding wife Sandra Serwah, for their support, understanding and immense contribution to my life and education.



ACKNOWLEDGMENT

No success can be achieved in isolation. This thesis would not have been possible without the strength, good health and sound mind from the Almighty God whom I believe as my Lord and personal savior. Truly, I owe it all unto the Lord. Unto Him, I give all the praise, the glory and the honour. Likewise, the success of this PhD thesis cannot be without the immense contribution of my supervisors. From the depth of my heart, I am indeed very grateful to my supervisors Dr. Yaw Marfo Missah, and Prof. Michael Asante for their immense, insightful contributions and deep thought-out suggestions towards this achievement. I also owe my late Uncle's wife Miss Elizabeth Frimpong, Dr. Iddrisu Wahab, Dr. Mark Amo-Boateng and Dr. Mary Antwi all at UENR and my siblings a lot of gratitude for their encouragement, support, suggestions and advice in undertaking this research. I would like to also express my profound gratitude to my Teaching Assistants, Pamela Akowua, Kojo Acheamong Nkatsia and Sarfo Manu Philip, my final year students, especially, Emmanuel Awoin and all the staff of the various banks for their immense role in the data collection. I also want to express my profound gratitude to all my colleague lectures at University of Energy and Natural Resources especially school of sciences lectures for all the valued support. God richly blesses you all. Finally, my thanks also go to Mr. Cosmas A. Rai of the department of Languages and General Studies, UENR, for proofreading this work.

ABSTRACT

Comparative to other methods, DEA is an improved method to organize and analyze data. However, it is very difficult to use only DEA to predict the efficiency and performance of other or new Decision Making Units (DMU). The main objective of this study is to build a high accuracy machine learning predictive models for predicting the efficiencies of banks by combining DEA with Machine Learning algorithm. The study built four Machine Learning Models namely; **DEA-DT**, **DEA-RF**, **DEA-NN** and **DEA-LR** to predict the efficiencies of banks. The study used 33% of the total bank branches in Ghana, largely in the nine regions. A two-stage DEA was used to determine the efficiencies of all bank branches and these banks were grouped based on a proposed algorithm, **Bank Classification Algorithm (BC Algorithm)**. In building the predictive models, 70% of the banks dataset were used to train and validate the models. The developed models were used to predict the efficiencies of the other 30% banks. A 10-fold Cross-Validation was applied to check the performance of all predicting models on each case dataset. All experiments were executed within a simulation environment and conducted in R studio using R programming language. Standardized Machine Learning evaluation metrics were used to compare the models. The results suggested a very good performance of all the machine learning models proposed by the study. However, a comparison among them clearly indicated a much better performance by the **DEA-RF** for predicting banks' efficiency in collecting deposit and **DEA-DT** for predicting banks' efficiency in investing deposits. This study has demonstrated that combining two models improve the performance, predictions and classification accuracies suggested by previous studies. In conclusion, the study proposed the usage of the proposed **BC Algorithm** for classifying banks based on their efficiencies in deposit stage and investment stage.

TABLE OF CONTENTS

DECLARATION	ii
--------------------------	-----------

DEDICATION	iii
ACKNOWLEDGMENT	v
ABSTRACT	vi
TABLE OF CONTENTS	vi
LIST OF FIGURES	xiii
LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE	xv
CHAPTER ONE.....	1
INTRODUCTION	1
1.0 BACKGROUND TO THE STUDY.....	1
1.1 PROBLEM STATEMENT	5
1.2 MAIN AND SPECIFIC OBJECTIVES OF THE STUDY.....	7
1.2.1 SPECIFIC OBJECTIVES	7
1.3 SIGNIFICANCE AND BENEFITS OF THE STUDY.....	8
1.4 SCOPE OF STUDY.....	8
1.5 ORGANISATION AND STRUCTURE OF THE THESIS	9
CHAPTER TWO.....	10
LITERATURE REVIEW	10
2.0 Introduction	10
2.1.2 The Data Structures of DEA	11
2.1.3 Data Envelopment Analysis (DEA) Algorithm.....	12
2.1.4 Advantages and Disadvantages of DEA	14
2.1.5 Application of Data Envelopment Analysis (DEA).....	15

2.2 OVERVIEW OF MACHINE LEARNING ALGORITHMS	20
2.2.1 Background	20
2.2.2 Types of Machine Learning	21
2.2.3 Topmost Machine Learning Algorithms	22
2.2.3.1 Logistic Regression	24
Types of Logistic Regression	25
Representation Used for Logistic Regression	25
Learning the Logistic Regression Model.....	26
2.2.3.2 Decision Tree	26
2.2.3.3 Random Forest.....	33
2.2.3.5 The Back-Propagation Algorithm	40
2.2.4 Evaluation of Machine Learning Algorithms Models.....	43
2.2.5 Application of Machine Learning Algorithms	48
2.3 APPLICATION OF COMBINED DEA AND MACHINE LEARNING ALGORITHM	53
2.4 CHAPTER SUMMARY.....	58
CHAPTER 3	61
METHODOLOGY	61
3.0 Introduction	61
3.1 DATA	61
3.1.1 Study Area and Data Description.....	61
3.1.2 Data Collection and Sample Size.....	63
3.2 THE PROPOSED BANK EFFICIENCY PREDICTION FRAMEWORK	64
3.2.1 Dataset for the Model Development	65

3.2.1.1 Determining the Response Variables (Bank Efficiency Scores and Classes)	65
3.2.1.1.2 Classification of the Banks' Efficiencies Scores	74
3.2.1.2 Predictor Variables.....	77
3.3 THE COMPUTER PROGRAMMING PLATFORM.....	77
3.4 BUILDING THE PREDICTIVE MODELS FOR PREDICTING BANKS EFFICIENCIES.....	77
3.4.1.1 The C5.0 Algorithm	77
3.4.2 The Random Forest (RF) Algorithm	80
3.4.3 The Artificial Neural Network	84
3.4.4 The Logistic Regression Algorithm	87
3.4.5 Metrics for Evaluating Machine Learning Algorithms	89
CHAPTER 4	91
RESULTS AND DISCUSSIONS.....	91
4.0 Introduction	91
4.1 DISCUSSIONS OF THE PREDICTION OF THE BANKS USING DECISION TREE ALGORITHM	91
4.2 DISCUSSIONS OF THE PREDICTION OF THE BANKS USING RANDOM FOREST ALGORITHM	93
4.3 DISCUSSIONS OF THE PREDICTION OF THE BANKS USING ARTIFICIAL NEURAL NETWORK	94
4.4 DISCUSSIONS OF THE BANKS' EFFICIENCY SCORES AND CLASSIFICATION	96

4.2.1 Bank Efficiency Determination Using the Adapted DEA Two-Phase Build Hull Algorithm	96
4.2.2 Using the proposed Banks' Classification Algorithm (BC Algorithm) to classify Banks Based on their Efficiencies Scores.....	102
4.5 DISCUSSION OF THE PROPOSED BANK EFFICIENCY PREDICTION FRAMEWORK	109
4.5.1 DISCUSSIONS OF THE PREDICTION OF THE BANKS EFFICIENCIES USING THE DEA-DT MODEL	110
4.5.1 The Bank branches (132 DMUs) efficiency analysis using the DEA-DT model Result	111
4.5.1.1 Case 2 –The Deposit Stage	111
4.5.1.2 Case 4-The Investment Stage	112
4.5.2 RESULTS AND DISCUSSIONS OF THE BANK BRANCHES (132 DMUS) EFFICIENCY ANALYSIS USING THE DEA-RF ALGORITHM PREDICTIVE MODEL	113
4.5.2 The Bank branches (132 DMUs) efficiency analysis using the RF model Result	113
4.5.2.1 Random Forest Order of Significant Predictor Variables	114
4.5.2.2 Case 2-The Deposit Stage.....	114
4.5.2.3 Case 4 –The Investment Stage	115
4.5.3 EMPIRICAL RESULTS AND DISCUSSIONS OF THE PREDICTIONS BY THE DEA-NN MODEL.....	117
4.5.3.1 The Bank Branches (132 DMUs) Efficiency Analysis Using the DEA-NN Model.	118
4.5.3.2 Case 2 –The Deposit Stage	118
4.5.3.3 Case 4 –The Investment Stage	119

4.5.4 EMPIRICAL RESULTS AND DISCUSSIONS OF THE PREDICTIONS BY THE DEA-LR MODEL.....	121
4.5.3.1 The Bank Branches (132 DMUs) Efficiency Analysis Using the DEA-LR Model..	122
4.5.3.2 Case 2 –The Deposit Stage	122
4.5.3.3 Case 4 –The Investment Stage	123
CHAPTER 5.....	125
GENERAL DISCUSSION	125
5.0 INTRODUCTION	125
5.1 ANALYSIS OF THE MACHINE LEARNING ALGORITHMS	126
5.3 Discussion	130
5.4 ORIGINALITY AND CONTRIBUTIONS TO KNOWLEDGE	134
CHAPTER 6.....	137
6.0 CONCLUSION AND RECOMMENDATIONS.....	137
6.1 LIMITATIONS OF THE FINDINGS.....	138
6.2 RECOMMENDATIONS	141
6.3 FUTURE STUDIES	142
LIST OF REFERENCES	143
APPENDICES.....	162

LIST OF TABLES

Table 4.1.: Efficiency Classes (case 1) (Authors construct).	104
Table 4.2: Efficiency classes (Case 2) (Authors construct).	105
Table 4.3: Efficiency classes (Case 3) (Authors construct).	107
Table 4.4: Efficiency classes (Case 4) (Authors construct).	108
Table 5.1: Performance of the Models Using Machine Learning Evaluation Metrics under the Deposit Stage	126
Table 5.2: Performance of the Models Using Machine Learning Evaluation Metrics under the Investment Stage.....	128

LIST OF FIGURES

Figure 2.1: A Graph of a Logistic Regression Sigmoid Function (Wickramasinghe & Karunasekara, 2016)	24
Figure 2.2: A Classic Example of Decision Tree (Shalev-Shwartz & Ben-david, 2014)	27
Figure 2.3: A Classic Example of Random Forest (Cutler et al., 2011).....	34
Figure 2.4: Classical NN model (Shamisi et al., 2011)	37
Figure 2.5: Single node in a MLP network source (Koivo, 2008)	38
Figure 2.6: A MLP network with one hidden layer source (Koivo, 2008)	39
Figure 2.7: A two-layer MPL Network	41
Figure 2.8 NN model with back -propagated errors	43
Figure 3.1: Map of the case study area showing the distribution of bank branches used for the Study (Author's construct).	62

Figure 3.2: The Banks efficiency prediction framework (Author’s construct)	65
Figure 3.3: The Classical two-stage DEA model taking two (2) inputs adapted from Wu (2006).	66
Figure 3.4: The Proposed Dual Role DEA Model adopted from Appiahene et al. (2019).....	71
Figure 3.5: Screen shot of R studio used for the programming (Author’s construct).....	73
Figure 3.6: The proposed framework (Author’s construct).....	86
Figure 3.7 : The Logistic Regression Steps (Donges, 2019)	88
Figure 4.1: A graph showing the deposit efficiency scores of the 444 DMUs (Author’s Construct)	97
Figure 4.2: A graph showing the investment efficiency scores of the 444 DMUs (Author’s Construct).	98
Figure 4.3: A graph showing the overall efficiency scores of the 444 DMUs (Author’s Construct)	99
Figure 4.4: A graph showing the efficiency scores of deposit, investment and overall stages of The 444 DMUs (Author’s construct)	101
Figure 4.5: A bar graph showing the proposed efficiency classes (Case 1) and the number of DMUs (Authors construct).	105
Figure 4.6: A bar graph showing the proposed efficiency classes (Case 2) and the number of DMUs (Authors construct).	106
Figure 4.7: A bar graph showing the proposed efficiency classes (Case 3) and the number of DMUs. (Authors construct).	107
Figure 4.8: A bar graph showing the proposed efficiency classes (Case 4) and the number of DMUs (Authors construct).	108
Figure 5.1: Graph showing the Performance of the Machine Learning Algorithms models and the four Models proposed in Case 2 (Author’s construct).	127

Figure 5.2: Graph showing the Performance of the Machine Learning Algorithms models and the four Models proposed in -Case 4 (Author's construct)..... 129

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

%PL	Rate of Performing Loans
AC	Overall accuracy
AI	Artificial Intelligent
ANN	Artificial Neural Network

AUC	
BCC	
BOG	Bank of Ghana
CAR	Capital Adequacy Ratio
CCR	Charnes, Cooper and Rhodes
CPU	Central Processing Unit
CRS	Constant Returns To Scale
CV	Cross-Validation
DEA	Data Envelopment Analysis
DEA-DT	Data Envelopment Analysis Decision Tree Model
DEA-LR	Data Envelopment Analysis Logistic Regression Model
DEA-NN	Data Envelopment Analysis Neural Network Model
DEA-RF	Data Envelopment Analysis Random Forest Model
DMU	Decision making Unit
DT	Decision Tree
EFA	Exploratory Factor Analysis
EFA	Exploratory factor analysis
ERP	Enterprise Resource Planning
FORTTRAN	Formula Translation
GDP	Gross Domestic Product
GDP	Gross domestic product
GPT	General Purpose Technologies
	International Conference on Applied Sciences and Technology
	International Conference on Data Mining
	Information and communication Technology
	Institute of Electrical and Electronic Engineering
	Area Under Curve
	Banker, Charnes and Cooper

ICAST

ICDM

ICT

IEEE

IMF International Monetary Fund

IS Information Systems

IT Information Technology

KNN	
LR	
MAE	
MAPE	Mean Absolute Percentage Error
MARS	Multivariate Adaptive Regression Splines
MARS	Multivariate Adaptive Regression Splines
MDA	Mean Decrease of Accuracy
MENA	Middle East and North African
MIS	Management Information System
MLP	Multilayer Layer Perceptron
MLR	Multinomial Logistic Regression
MRA	Multinomial Regression Analysis
MRA	Multinomial Regression Analysis
MSE	Mean Squared Error
NA	Not Available
NAICS	North America Industry Classification System
NN	Neural Networks
OCAM	Office, Computing and Accounting Machinery
PCA	Principal Component Analysis
PCBS	Palestinian Central Bureau of Statistics
	Performing Loans
	probabilistic neural network
	Random Access Memory
	Relative Bias
	K-Nearest Neighbors Logistic
	Regression
	Mean Absolute Error

PL

PNN RAM

 rBIAS

RF Random Forests

RF Random Forest

RF Random Forest

RMSE	
RMSPE	Root Mean Square Percentage Error
ROC	Receiver Operating Characteristic
RST	Rough Set Theory
RWA	Risk-Weighted Assets
SBU	Strategic Business Unit
SFMC	Smoke-Free Melaka Campaign
SIC	State Insurance Corporation
SME	Small and Medium Scale Enterprise
SOFM	Self-Organizing Feature Map
SSNIT	Social Security and National Insurance Trust
SVM	Support Vector Machines
UENR	University of Energy and Natural Resources
VRS	Variable Returns to Scale
VRS	Variable Returns to Scale
VRS	Variable Return to Scale
WACB	West African Currency and divided the Board

Root Mean
Square Error

CHAPTER ONE

INTRODUCTION

1.0 BACKGROUND TO THE STUDY

A significant growth in performance and efficiency cements the basis for improvement in the standard of living (Majeed & Ayub, 2018 and Niebel & Mannheim, 2014). The prosperity of any country in terms of its economic fortunes is ultimately based on performance and efficiency of its important institutions like the banks (Cardona, Kretschmer, & Strobel, 2013; Leung & Zhang, 2016 and Stanley, Doucouliagos, & Steel, 2018). This means that the survival and continuous existence of banks within the financial industry is of great importance to every economy and all of its citizens (Antonija et al., 2017; Navapan, Liu, & Sriboonchitta, 2017 and Sahoo, 2014). As far as financial institutions such as banks play a significant role of financial mediations, the determination of their efficiency is essential and needs to be given enough attention. The significance of the banking industry is also premised on the ground that these financial institutions happen to be key channels of investments and allocations of credit in an economy (Sufian et al., 2016). The banking industry offers vital financial intermediation role by transforming deposits into prolific investments. Unlike countries like the USA where financial markets and the banking industry works in harmony to channel capitals, in countries like Ghana, financial markets are stunted and sometimes absolutely absent (Sufian et al., 2016). It is therefore the responsibility of the banking industry to close the gap between investors and debtors. The banking industry controls most of the financial flows and accounts of the financial system's total assets (Sufian et al., 2016). Therefore, it is prudent to presume that an efficient banking industry may help ensure an effective financial system which is conducive to economic growth and development.

This concept of efficiency is an old one and also quantifiable as it can be calculated through the assessment of the useful output ratio to total input. The traditional efficiency measurement of bank performance using the Data Envelopment Analysis (DEA) model usually accounts for

input and desirable output only (Hamid et al., 2017). According to a report by the Development & Research Center in 2004, firms that effectively utilize their resources in their business process and operations experience greater efficiency and performance. This would lead them to greater competitiveness that promotes sustainable economic growth. Most banks and their management have little or no idea on how to predict the real efficiency of their organizations. Even though assessing the concept of efficiency of banks has been extensively studied, a study by Burki & Dashti (2003) cited by (Hamid et al., 2017) assessed the cost efficiency of Kuwaiti banks and suggested the cost inefficiency was found within the Kuwait banks. This according to the researchers was caused by locative inefficiency while the score for technical inefficiency was very high (Hamid et al., 2017). The analysis of the study was based on traditional Data Envelopment Analysis (DEA) measures of organizations' performance (Hamid et al., 2017). Thus efficiency as its best suits the requirements of management (Lefley, 2015; Wong & Dow, 2011). A DEA is a linear-programming-based method which is used to determine the comparative efficiency of homogenous organizational units such as banks, school, governmental and non-governmental agencies, tax offices etc. A DEA model as a nonparametric technique has been used as a single method (Alexander et al., 2007; Álvarez et al., 2016; Necmi, 2006; Cao & Yang, 2011 and Paço & Pèrez, 2015) or combined (Alinezhad, 2016; Anthanassopoulos & Curram, 1996; Azadeh et al., 2011; Lee, 2010; Razavi et al., 2013 and Wu et al., 2006) with others extensively in previous literatures to measure the efficiency of an organization. Data Envelopment Analysis (DEA) has also been suggested to be a good qualitative measure of organizational efficiency. For instance, Cao & Yang (2011); Chen et al. (2006); Madjid et al. (2009); Wang et al. (1997); Hatefi & Fasanghari (2014) and Appiahene et al. (2019) applied DEA to assess the impact of IT on the performance of firms and concluded a positive impact. According to Aggelopoulos & Georgopoulos (2017); Wanke et al. (2016), compared to other methods, DEA is superior method to analyze and ascertain

productivity since it enables effectiveness to change after some time. It also requires no earlier supposition on the detail of the best practice frontier.

Traditionally, DEA has been one of the most popular tools to assess a firm's performance. Machine Learning Algorithms have also been used for the prediction of bank efficiency and even its bankruptcy and such studies include (Chang et al., 1996; Dash et al., 2006; Jardin , 2018 and Mai et al., 2018). Neural Networks (NN), for instance have recognized their role as data analysis tools in different areas. Decision tree algorithm as used (Chen, 2016; Santos et al., 2017 and Wu, 2006) and Random Forest algorithms have also been found to be an efficient prediction method. Various techniques and machine learning algorithms have been applied in prediction studies but combining DEA with Machine learning algorithms to predict the efficiency of banks is really scarce especially using data from developing country.

Comparison of bank branches' efficiency across developing countries are also completely lacking in literature (Mohd, 2001 and Tra et al., 2018).

This study differs from previous studies (Hamad & Anouze ,2015; Kwon & Lee, 2015 and Wu 2006) in the following aspects: first and foremost, this study combines a two-stage DEA model with different Machine Learning Algorithms namely Decision Tree, Random Forest, Artificial Neural Network and Logistic Regression, where the efficiency at each stage (thus efficiency in collecting deposit from customers, Deposit efficiency and efficiency in investing the deposit, Investment efficiency) is also considered as input for the next stage using a huge dataset a from a developing country. The study also classified the various banks based on their efficiency scores using a proposed algorithm, Bank Classification Algorithm suggested by the study. This efficiency classes (Class1, Class2 Class 3 and Class 4) were used as the response variable making, the response variable for the study categorical. For the development of the four models, the study considered predictor variables which were both internal and external and directly influence bank performance and efficiency as suggested by (Thanassoulis et al., 2008).

The development and performance of the proposed in this current study was done on big dataset compared to Wu (2006) work which was done on a small dataset from existing literature (Wang et al., 1997). Kwon & Lee (2015) also used a total of 181 DMUs as a dataset. Which accooidng to this same Kwon & Lee (2015), a large data set is often favored for the generalization of models. A K-fold cross-validation was used to better the performnace of the models as compared to Wu (2006) works, which was based on V-fold cross validation. The data used in this study is not a cross-country bank-level data compared to Hamad & Anouze (2015). In this study ,70% of the data set (444 DMUs) was used for training and 10 % validation and 30% for testing the models compared to Kwon & Lee, (2015) study of combining DEA with Back Propagation Neural Network which was based on a ratio of 60:20:20 for training,validating and testing respectively. This is also a single study that has build and proposed four different models that has yieded fovorable classification accuracy rate. These models were developed by combining DEA with Decision Tree (DEADT), Random Forest (DEA-RF), Artificial Neural Network (DEA-NN) and Logistic Regression (DEA-LR). In the case of the DEA-DT, the model perfomed better than Wu (2006) work by giving an accuracy of not less than 90% compreaed to Wu (2006) 69.44%. Comparing the DEA-RF to that of Hamad & Anouze (2015), the study also suggested a fovuorable classification accuracy of 90% compared to Hamad & Anouze (2015) work that also yieded 79.4% and finally, the DEA-NN suggeseted a favouable Mean Absolute Percentage Error (MAPE) of about 24.6% compared the BPNN MAPE of 36.9% also suggsetd by (Kwon & Lee, 2015). Compared to the works of Hamad & Anouze (2015); Kwon & Lee, (2015) and Wu (2006), this study demonstrated the performance of the four proposed models by using five standard machine learning evaluation metrics namely; Root Mean Square Error(RMSE), Mean Absolute Percentage Error(MAPE), Mean Absolute Error(MAE), Root Mean Square Percentage Error (RMSPE), and rBIAS.

1.1 PROBLEM STATEMENT

The assessment of banks efficiencies creates serious problems for banks and their managers (Adusei, 2016; Sufian et al., 2016 and Titko, Stankevičienė, & Lāce, 2014). Due to the critical significant role of banks in the national economy, their efficiencies and performances are a hotly debated topic in the academic and business environments (Titko et al., 2014). Models and frameworks designed for this assessment are normally based on econometric analysis (Brynjolfsson, 1993; Dedrick et al., 2013; Dedrick et al., 2003; Kılıçaslan et al., 2017; Nations, 2008 and Roghieh et al., 2004) and other parametric methods. Traditionally, financial institutions such as banks have concentrated on numerous profitability measures as a means to assess their efficiency and even performance by using multiple ratios based on different aspects of the operations (Dash et al., 2006; Wanke et al., 2016). Nevertheless, ratio analysis suggests comparatively insignificant amount of information when considering the approximation of overall efficiency measures of firms (Tavana et al., 2016). As an option to this traditional method of efficiency assessment, frontier efficiency analysis allows one to empirically recognize best practices in operational environments.

Non-parametric techniques such as Data Envelopment Analysis (DEA) have been suggested in literature as technique to evaluate bank efficiency (Lampe & Hilgers, 2015). Compared to other methods, DEA is an improved method to organize and analyze data since it allows efficiency to change over time (Lampe & Hilgers, 2014). It also does not involve any previous postulation on the bases of best practice frontier (Aggelopoulos & Georgopoulos, 2017; Fallahpour et al., 2017 and LaPlante & Paradi, 2014). However, such studies (Chen et al., 2006; Chen & Zhu, 2004a; Izadikhah et al., 2017; Madjid et al., 2009; Tavana et al., 2017 and Wang et al., 1997) that used only non-parametric methods such as pure DEA and its models suffer from weak discrimination power (Ashoor, 2012; Chen, 2016; Santos et al., 2017 and Wu, 2006, 2009) and also sensitive to the presence of outliers and statistical noise (Da et al., 2018 and Dash et al., 2006). It is also very difficult to use only DEA to predict the efficiency and performance of other or new Decision Making (Emrouznejad & Yang, 2017;

LaPlante & Paradi, 2014 and Wanke et al., 2016). Literature on DEA and Machine Learning Algorithms using data from Ghanaian banks is also scarce.

To overcome this problem, more promising and novel methods and procedures for the prediction and classification of bank efficiencies are necessary (Sreedhara et al., 2018). Combined DEA and Machine Learning Algorithms can offer these suitable and scientific methods. A lot of machine learning algorithms have been developed and utilized in the business and financial sectors for predictions and forecasting. For instance, Chen & Hao (2017) used Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for stock market indices prediction using the two well-known Chinese stock market indices. There are other studies such as Abellán & Castellano (2017); Ahn et al. (2000); Anandarajan et al. (2001); Caggiano et al. (2014); Caggiano et al. (2016); Göçken et al. (2016); Janek et al. (2016); Kim & Han (2000); Patel et al. (2014b, 2014a); Qiu et al. (2016) and Tsai & Wu, (2008) that have implemented machine learning algorithms in the business and financial sectors in forecasting and prediction studies.

According to literature, Machine Learning Algorithms used to build predictive models have been suggested as the one of finest and best predicting methods with a high accuracy and validity in the field business and financial forecasting (Göçken et al., 2016). Nevertheless, a combined high performance machine learning model for prediction and classification of bank branches' using their efficiency across developing countries are also completely lacking in literature (Mohd, 2001 and Tra et al., 2018) especially using a combined DEA and Machine Learning Algorithms. The results of models built in these studies were also weak in terms of performance (Yang et al., 2015). Combining two or more models has gained a lot of attentions and interest across various disciplines comprising Management and Decision Sciences, Applied Mathematics, Data Science, Machine Learning and Artificial Intelligent (Yang et al., 2015 and Zheng et al., 2004). According to Yang et al. (2015) and Zheng et al.

(2004), combined models consistently outperform individual models for classification and prediction tasks. This current work uses a primary data gathered from more than four hundred (400) DMUs of bank branches.

1.2 MAIN AND SPECIFIC OBJECTIVES OF THE STUDY

The main objective of this study is to build a strong and robust predictive model for predicting the efficiency of banks by combining DEA with Machine Learning Algorithms using data from the Ghanaian banking sector as a case study. The output of this study is expected to be predictive models that can be used to predict the efficiency of Ghanaian banks accurately. This main objective was realized through the following specific objectives:

1.2.1 SPECIFIC OBJECTIVES

1. To use a Decision Tree algorithm to build a model for predicting bank efficiencies using dataset from the Ghanaian banking sector.
2. To use a Random Forest algorithm to build a model for predicting bank efficiencies using dataset from the Ghanaian banking sector.
3. To use an Artificial Neural Network to build a model for predicting bank efficiencies using dataset from the Ghanaian banking sector.
4. To use a Logistic Regression Algorithm to build a model for predicting bank efficiencies using dataset from the Ghanaian banking sector.
5. To adapt an existing algorithm to determine the efficiencies of the selected banks and use an algorithm to classify the banks based on their efficiencies in taking deposits and investing the deposits.
6. To develop and propose a high performance models for predicting bank efficiencies by combining DEA with: Decision Tree Algorithm (DEA-DT), Random Forest Algorithm (DEA-RF), Artificial Neural Network (DEA-NN) and Logistic Regression Algorithm (DEA-LR).

1.3 SIGNIFICANCE AND BENEFITS OF THE STUDY

The significance and benefits of this study include but not limited to the following:

1. Researchers can also use the proposed framework and models in the study to achieve high prediction accuracies in their studies.
2. The general Ghanaian banking industry and other firms may benefit from the results of this thesis as they may have a better understanding of relationship between resources and banks performance and efficiency.
3. Customers' and investors may benefit from the findings of this study as the finding would add value to their knowledge of how efficient their various bank branches are in terms of managing their capitals.
4. The predictive models build in the study can be used to assess and predict the efficiency of new banks going forward.
5. The proposed Bank Classification Algorithm (BC Algorithm) by the study can be used to classify banks in Ghana based on their efficiency in both deposit stage and investment stage.

1.4 SCOPE OF STUDY

The study was undertaken basically to classify and predict the efficiencies of banks by combining DEA with Machine Learning Algorithms. The case study firms were commercial bank branches in Ghana. Due to the large number of universal bank branches operating in the country as of the time of this study (2016), the study was limited to 33% (comprising 17 universal banks), of the total bank branches in Ghana mostly in Greater Accra, Ashanti, Western, Brong Ahafo, Eastern, Northern, Upper East, Volta and Central Regions. Part of the data for the study was also audited 2016 financial statements from the various banks. With respect to factors that affect bank's deposit collection, there were other several factors that were not taken into consideration because of inadequate data on these factors. This study is therefore limited to Bank's Fixed Asset, IT expenditure and Total Number of Employees as factors that can impact deposit mobilization by Banks.

1.5 ORGANISATION AND STRUCTURE OF THE THESIS

This thesis consist of six different chapters which have been structured in a logical sequence starting from the introductory chapter containing the background information of the topic, the problem statement, the objectives, the study's significance and scope . Chapter One is followed immediately by Chapter Two; literature review which also discusses the explanation and description of some concepts used in the study. It also contains information on extensive literature review of the topic, related works and finally conclusions drawn from these literature reviews. The Chapter Three which discusses the methodology describes comprehensively the various scientific methods, techniques and tools used to accomplish the study. The Chapter Four also deals with the presentation and discussions of the study results taking into consideration the five specific objectives.

Chapter Five also presents the general discussions and analysis of the study results. Finally the last chapter which is Chapter Six contains the conclusion drawn from the study, recommendations to various stakeholders who can use the study results and suggestions for future studies.

CHAPTER TWO

LITERATURE REVIEW

2.0 Introduction

The purpose of this study is to have a better understanding of the banks' efficiency measurement and how to combine machine learning algorithms with Data Envelopment Analysis (DEA) to build a high accuracy models for predicting the efficiency of banks and also classify these banks using their efficiencies. To provide the necessary background, this chapter will review literature on machine learning models for predicting banks' efficiency, DEA for assessing banks' efficiency, and previous applications of combined DEA and machine learning

algorithms models for predictions. The review is done in a way as to identify gaps in existing literature and present how this thesis will address these gaps.

2.1 OVERVIEW OF DATA ENVELOPMENT ANALYSIS (DEA)

The fundamental efficiency is a proportion of output over input. To enhance efficiency, one needs to either increase the output or reduces the input. If there is an increase in both cases, the rate of increment for output ought to be more prominent than the rate of increment for inputs. On the other hand, if the two are diminishing, the rate of abatement for outputs has to be lower than the rate of reduction for inputs. Another approach to accomplish higher efficiency is to present innovative technologies like IT or to reengineer business process-lean administration which may decrease sources of inputs or increase the capacity to deliver more outputs (Adusei, 2016).

Data Envelopment Analysis (DEA) models can create new other options to enhance efficiency compared with different techniques. Linear programming is the backbone procedure that is based on enhancement platforms. Henceforth, what separates the DEA from different techniques is that it distinguishes the ideal methods for performance as opposed to the mean. Distinguishing proof of ideal performance prompts benchmarking in a normative way. Utilizing DEA, bank managers can recognize top performers, as well as find alternative approaches to goad banking firms into getting to be extraordinary compared to other banks. Since the fundamental work of Charnes, Cooper, & Rhodes (1978), DEA has been responsible for endless research literature within both the non-benefit and revenue driven sectors. Not long ago, the utilization of DEA within the banking sectors has been restricted to conferences and journals. Thus, bank managers have not accepted DEA as a standard instrument for benchmarking and decision making. This is attributed to complications in formulation and the disappointment of DEA experts to adequately close the theory gap. DEA is a linear approach

for distinguishing performance or its parts by considering various assets that are utilized to accomplish output or results in the banking sector (Peter Wanke et al., 2015).

These assessments can be done not just at the firm's level, yet additionally in subunits, for example, departmental comparisons, where numerous areas of improvement in saving specific input assets or techniques to expand the output can be recognized. In short, according to Ascarya et al. (2008), DEA can enable bank directors to achieve the following:

- Review their banks relative efficiency.
- Recognize top performance in firms advertise, and
- Recognize approaches to enhance their efficiency, if their bank is not one of the best performing banks.

2.1.2 The Data Structures of DEA

Data Envelopment Analysis determines productive and economic performance of a set $j = 1, 2, \dots, n$ observed DMUs and in our case banks in Ghana. These observations transform a vectors of $i = 1, 2, \dots, m$ inputs $x \in \mathbb{R}^{m++}$ into a vector of $i = 1, 2, \dots, s$ outputs $y \in \mathbb{R}^{s+}$ using the technology represented by the following constant returns to scale (CRS) proposed by Charnes et al., (1978) productivity possibilities set: $P_{CRS} = \{(x, y) \mid x \geq X \lambda, y \leq Y \lambda, \lambda \geq 0\}$, where $X = (x)_j \in \mathbb{R}^{s \times n}$, $Y = (y)_j \in \mathbb{R}^{m \times n++}$ and $\lambda = (\lambda_1, \dots, \lambda_n)$ is a semi-positive vector.

Data are considered as regular vectors and matrices, which makes up the inputs of the approximation functions. The set of the approximation functions produce a structure that contains field with the approximation output as well as the input of the approximation function (Álvarez et al., 2016).

We can directly access the field and the complete structure can also serve as an input to other functions that output results. According to Álvarez et al. (2016), the following are some of the filed structures:

- X , Y and Y_u : contain the inputs, outputs and undesirable outputs variables, respectively.

- n and $neval$: number of DMUs, and number of evaluated DMUs.
- m , s and r : number of inputs, outputs and undesirable outputs.
- $model$, $orient$, rts : strings containing the model type, the orientation, and the returns to scale assumption.
- eff : computed efficiency measure.
- $slackX$, $slackY$, $slackYu$: computed input, output and undesirable output slacks.
- $names$: names of the DMUs.

2.1.3 Data Envelopment Analysis (DEA) Algorithm

The traditional strategy for applying Data Envelopment Analysis (DEA) to dataset is repeatedly solving as many Linear Programmes (LPs) as there are entities (Dulá, 2011). The measurements of these LPs are typically dictated by the amount of dataset, and they retain their magnitudes as each decision-making unit is scored. This approach can be computationally difficult, particularly with huge dataset. To avoid this, Dulá (2011) proposed an algorithm based on a two-stage phase procedure where the first stage finds the extreme efficient entities, the frame of the production possibility set. The frame is then used in the second stage to score the rest of the entities. The new technique applies to any of the four standard DEA returns to scale. It likewise confers adaptability to a DEA study on the grounds that it puts off the choice about introduction, benchmarking estimations, and so forth, until after the frame has been identified. Comprehensive computational testing on large data sets confirms and authenticates the algorithm that it is computationally efficient.

Algorithm 2.1 : Two-Phase Algorithm BuildHull

[PHASE 1: FRAME IDENTIFICATION]

Input. \mathcal{A} \. The DEA data set.\.

Output. \mathcal{F} : \. The Frame of the DEA data set.\.

Step 0. Initialization.

1. $\mathcal{R} \leftarrow \mathcal{A}$, $l \leftarrow 1$ \. \mathcal{R} is a temporary work-space array. \.
2. Find one generator from \mathcal{R} that is an extreme element of the DEA hull; remove it from \mathcal{R} and place

it in \mathcal{A}^l

Step 1. Iteration. While $\mathfrak{R} \neq \emptyset$ Do:

1. Select some $\alpha' \in \mathfrak{R}$.

2. Set $\mathbf{b} \leftarrow \alpha'$ and solve LPs $P\mathbf{t}^{\text{CRS}}/D\mathbf{t}^{\text{CRS}}$.

$(\pi^*, \beta^*) \leftarrow$ Optimal dual solution.

3. If $(\pi^*, \mathbf{b}) + \beta^* > 0$ Then:

a. Find $a^* \in \mathfrak{R}$ such that:

$$\alpha^* = \arg \max_{\{a_j \in \mathfrak{R} \mid (\pi^*, a_j) + \beta^* > 0\}} (\pi^*, a_j) + \beta^*.$$

In the CRS case use Result 4B to find a^* . In case of a tie, set a^* to one extreme point among generators in support set.

b. $\mathfrak{R} \leftarrow \mathfrak{R} \setminus a^*$; $\mathcal{A}^l \leftarrow a^*$.

c. $l \leftarrow l + 1$.

Else $(\pi^*, \mathbf{b}) + \beta^* \leq 0$:

a. $\mathfrak{R} \leftarrow \mathfrak{R} \setminus a^*$.

End Do. Step 2. Conclusion. $\mathcal{F} \leftarrow \mathcal{A}^l$. End Phase 1.

[PHASE 2. SCORE DMUS.]

Classify and score all points in $\mathcal{A} \setminus \mathcal{F}$ using appropriate DEA LP.

End.

Mehrabiana (2013) also proposed another DEA algorithm for classification of DMUs efficient and inefficient units. Mehrabiana (2013) algorithm depended on non -Archimedean Charnes-Cooper-Rhodes (CCR) framework. The model also applies affirmation value for the non -Archimedean utilizing just basic calculations on inputs and outputs of DMUs.

The merging and effectiveness of the new algorithm demonstrate the benefit of this algorithm contrasted with the Thrall's algorithm.

2.1.4 Advantages and Disadvantages of DEA

2.1.4.1 Advantages of DEA

DEA is a model for analysis which obliges an exhaustive perspective of a firm's efficiency. It

is a good and standard tool for assessing the efficiency of a firm. This is partly because of the way

that a large number of subjective elements influence the quality and efficiency of the provision that should be managed well. In spite of the fact that there is no reasonable connection between inputs (expended) and output (created) in DEA, Ashoor (2012) recognized the following:

- Each DMU can be described exclusively.
- Inefficient DMUs are enhanced by projecting them on the efficient frontier (envelopment).
- DMU encourages influencing inductions for each DMU among the surmising on the DMUs' general to profile.
- Different input and numerous outputs can be dealt with in different DMUs estimations.
- An attention on a practice frontier, rather than on focal propensities.
- No confinements are forced on the useful frame relating inputs to outputs
- A focus on a best-practice frontier, instead of on central-tendencies

2.1.4.2 Disadvantages of DEA

Although conventional DEA models are viewed as a ground-breaking apparatus for productivity appraisal, numerous impediments have been distinguished:

- As DEA is an extraordinary point strategy, it is exceptionally sensitive to noise (notwithstanding for symmetrical noise with a zero mean) that may cause critical mistakes in effectiveness estimations.
- Statistical hypothesis tests are troublesome in light of the fact that DEA is a nonparametric.
- The standard definition of DEA depends on isolated linear programs for each DMU, which is computationally difficult.
- Cold-heartedness to impalpable and downright parts (e.g. benefit quality in a bank office).
- Troubles in accumulating diverse parts of effectiveness particularly at whatever point DMUs performs multiple activities.
- Cold-heartedness to impalpable and downright parts (e.g. benefit quality in bank offices).
- There are crucial problems related to mixing multiple dimensions in the analyses
- There are critical issues identified with blending numerous dimensions in the analyses.

- It is difficult to rank effective units totally on the grounds that all efficient DMUs should have efficiency score of 100% score.
- There is no precisely vigorous methodology for assessing or testing the suitability of a set of influences in an efficiency study.

2.1.5 Application of Data Envelopment Analysis (DEA)

Data envelopment analysis (DEA) is a non-parametric technique that creates a relative ratio of weighted outputs to inputs for each decision making unit i.e. a relative efficiency score (Avkiran, 2006). DEA attempts to address some of the explicit flaws of the growth accounting approach (LaPlante & Paradi, 2014). By enveloping the observed input–output combinations , DEA attains an estimate of the production frontier and uses it to detect the role of technological revolution productivity growth (Kılıçaslan et al., 2017). Data envelopment analysis is one of the topmost techniques used in efficiency measurement of firms (Titko et al., 2014). DEA has so many application areas but according to Paradi, Vela, & Zhu, (2010), bank branch studies have been and will continue to be one of the most predominant application areas of DEA. Current DEA application studies in banks include that of Sharif et al. (2019); Silva et al. (2018); Stewart et al. (2015); Vidyarthi (2018); Wang et al. (2019); Wang et al. (2018); Zha et al. (2016) and Zhou et al. (2018).

As noted by LaPlante & Paradi (2014), little attention have been given to assessing the growth potential of individual bank branches. Based on this, the authors presented five different models for assessing the efficiencies of branch network of one of Canada’s top five banks using DEA. The study suggested that two of the proposed models were able to successfully identified best performing branches using their efficiency scores. Avkiran (2014) applied a Dynamic Network Data Envelopment Analysis (DN- DEA) using a commercial bank as a case study with emphasis on testing the banks heftiness. The author used 16 external banks in China as benchmarked against 32 local banks for the post-2007 era that follows major reforms. The results of the study which was an illustrative one suggested there was no statistical

significant difference between the Chinese local and foreign banks performance based on mean overall efficiency estimates. They anticipated the need for progressively more sophisticated analysis tools such as DN-DEA in order to explore extensively bank performance.

Fukuyama & Weber (2014) measured the Japanese bank performance using a dynamic Network DEA approach and suggested that for a 3 year dynamic window, inefficiency in banks ranges from 19.5 % of average outputs and inputs in 2007–2009 to 21.5 % of average outputs and inputs in 2008–2010. The authors indicated that a lot of banks in the data sample can improve their efficiency by collecting more deposits from customers in the first stage and then using the deposits to generate a larger collection of loans and securities in the second stage of operation (Fukuyama & Weber, 2014).

Similar to Fukuyama & Weber (2014), Duygun et al. (2015) also used a Network Data Envelopment Analysis approach which consisted of two stages to assess the efficiency of various European Airlines and suggested that most of the inefficiencies were generated during the first stage of the analysis. Kaffash et al. (2017) in a study proposed a new version of the modified semi-oriented radial DEA measure. To illustrate their proposed model they employed two widely used selections of inputs and outputs to estimate the efficiency scores for a sample of banks operating in Persian Gulf Council Countries (GCC) over the period of 2002–2011. The finding shows that banks operating in environment with a relatively lower risk of a banking crisis were more efficient (Kaffash et al., 2017). Empirical results by (Muhammad et al., 2018) also revealed a mix trend among Saudi banks 2008-2016 in achieving technical, pure technical and scale efficiency. As the Banking industry plays a critical role in the economic development of a country (Hamid et al., 2017). Hamid et al. (2017) applied DEA to measure the efficiency of the banking sector in the presence of Nonperforming Loan in Malaysia and suggested that, modelling the efficiency in the absence of objectionable outputs can give misleading results and biased assessment. Hamid et al. (2017) compared both results between the domestic and

foreign banks and suggested that the DEA technical efficiency score for domestic banks was slightly greater than the foreign banks.

Havidz & Setiawan (2015) investigated the efficiency of Indonesian Islamic Banks by employing Data Envelopment Analysis (DEA) approach. The authors' data covered the periods of January 2008 – September 2014 which was based on the quarterly-published report from Indonesian Central Bank. The findings of their study show that none of the banks was consistently efficient for all periods of research by Overall Technical Efficiency (OTE), Pure Technical Efficiency (PTE), and Scale Efficiency (SE). The overall results also suggest that efficiency of Banks was significantly affected by Return On Assets (ROA), Operational Efficiency Ratio (OER), and Inflation Rates (INF), while Financing to Deposit Ratio (FDR), Capital Adequacy Ratio (CAR), size, and GDP growth rate have insignificant effect on bank efficiency. Titko et al. (2014) contributed to the existing analytical data on bank performance in Latvia by applying DEA to measure the efficiency of selected banks in Latvia. Based on their study, the authors' developed fourteen alternative model specifications by using the results of earlier conducted correlation analysis.

The growing investments and application of IT in organizations using various models, techniques and methods to achieve competitive advantage has generated the IT impact productivity debate popular called "Productivity Paradox" (Han et al., 2011). Several scholars, Brynjolfsson (1996); Brynjolfsson (1993) and Ko & Osei-Bryson (2004) in an attempt to address the measurement issue of IT impact on productivity either used only DEA, Cao & Yang (2013); Chen et al. (2006); Chen & Zhu (2004); Madjid et al. (2009); Sigala 2003 and Wang et al. (1997) or combined it with other methods, (Ko & Osei-Bryson, 2006; Ngai et al., 2009; Shao & Lin, 2001; Wu, 2009).

For example, Paço & Pérez (2015) attempted to evaluate the impact of ICT on the productivity of hotels in Portugal through Data Envelopment Analysis (DEA). The study did not only demonstrated how important ICT is in realizing advanced levels of productivity but also

discussed other explicit concerns which should be taken into consideration so that the positive returns of the investment in ICT can be achieved. Paço & Pèrez (2015) found that the accessibility of ICT does not alone lead to optimal performance. Using data from Iranian manufacturing businesses during 2002–2006, Abri & Mahmoudzadeh (2014) applied the method DEA to study the subject matter. Results show that IT has positive and significant impact on the productivity of manufacturing companies. This positive impact would be experienced in high IT-intensive businesses more than the others. Wang et al. (1997) through their DEA model and methodology also evaluated the marginal benefits of IT using 36 DMUs financial institutions data. The study suggested that for a collection of IT investment, IT impacted substantially on organizations revenues but they were quick to contest their own results due to their beliefs that there was no relapse between IT investment and profits to performance and that IT was exclusively used in stage one. Using twenty Iran, conventional power plants data, Madjid et al. (2009) assessed the IT impact on productivity in conventional power plants and suggested a DEA model that permits the incorporation of production performance and investment performance.

Chen & Zhu (2004) also used DEA model on a 27 banks. They considered IT investment, fixed asset and number of employees as inputs in stage I that can be utilized to collect funds in a form of deposits from customers. The outputs of this two-stage DEA were also profit and fractions of loans recovered from customers. Chen et al. (2006) used 27 DMUs of banks suggested only three firms units as efficient in the two efficient calculation phases. Their results suggest that the bank's investment in IT, assets and employee should be assigned to only one specific stage. Chen et al. (2006) used a two-stage DEA model with three inputs (IT investment, fixed assets and number of Employees) and overall output as profit generated from investing the deposit from customers in securities and also given out loans to customers and fractions of loans recovered as was done by (Chen & Zhu, 2004 and Wang et al., 1997).

Finally, Sheng et al.,(2002), also employed DEA to propose a framework that can be used to measure the productivity of organizational IT investment.

2.2 OVERVIEW OF MACHINE LEARNING ALGORITHMS

2.2.1 Background

In order to deal with a task on a computer, we must have a series of commands that should be followed to process the data to information popularly known as computer algorithm. For example, one can design a bubble sort algorithm for sorting numbers. The numbers in this case are going to be the inputs while the preferred ordered list would be the output. There can be many different algorithms to be designed to do the same work but computer scientist would be much interested in efficient algorithms that use less memory and space. On the other hand, there are situations that do not follow normal algorithm design. For instance, to sort email messages into authentic and spam, the input is an email text that in the simplest form is a file of characters.

The output is either YES/NO to specify whether the message is spam or authentic. Transforming input to output in such scenarios is very difficult because, what would be spam in email varies from user to other. So to deal with such situations, one can use the data available by simply compiling large volumes of emails dataset where we know some of these dataset to be spam easily. We can therefore “learn” the characteristics or feature that makes an email spam from the spam sample in the dataset. Thus, the computer called “machine” would be used to automatically extract its own algorithms (model) for this kind of task. According to Omary & Mtenzi (2010) computer is a machine for executing or aiding calculation; it receives data, process them and output information based on an algorithm on how the data has to be process the data.

Whenever there is design of an algorithm for a computer task such as sorting, there is no need to spend time to lean sorting of numbers (Alpaydın, 2010). Nevertheless there are many other applications that designing an algorithm for all would be difficult or not even possible. In this

case, we can construct a good and important approximation even though this explanation may not take care of all the data but one can be sure of accounting for some part of the dataset. This is achieved by detecting certain patterns or characteristics and using same for new dataset. This method or process is termed as Machine Learning and has so many applications retail, finance, manufacturing, medicine, telecommunication, agriculture, sports, etc.

When Machine Learning is applied to large dataset, it becomes data mining but is not just about database; but also part of artificial intelligence (Alpaydm, 2010). Machine learning is, therefore, about programming computers (Machines) to optimize a performance criterion using example data or past experience (Alpaydm, 2010). We develop a model well-defined to some few parameters, and learning is the operation of a computer programme to advance the parameters of the model utilizing the training data or past experience. The model can be predictive to make predictions in the future, or descriptive to gain knowledge from data, or both. Machine learning utilizes the hypothesis of statistics in building mathematical models, in light of the fact that the core mandate is making induction from dataset sample (Alpaydm, 2010).

2.2.2 Types of Machine Learning

According to Shalev-Shwartz & Ben-david (2014), there are three (3) different types of Machine Learning discussed as follows:

- **Supervised Learning:** As the name implies, this type of machine learning algorithm is done under the supervision or assistance of a supervisor or teacher (Alpaydm, 2010). Thus, the entire learning process depends on the teacher or under the care of the teacher.

The algorithm is made up of dataset which contains both response and predictor variables (Smola & Vishwanathan, 2008). We, therefore, build a function that processes the inputs to our desired output. There is a training process that proceeds until the point when the model accomplishes a coveted level of exactness on the training data which is normally percentage of the entire dataset. Examples are Regression, Decision Tree, Random Forest,

KNN, Logistic Regression, Neural Network etc. (Breiman, 2001a; Mashat et al., 2012; Omary & Mtenzi, 2010 and Scornet, 2010).

- **Unsupervised Learning:** Unlike the previous supervised algorithms, these types do not have or contain any response or target variable of interest to predict (Alpaydm, 2010). This type of learning is independent and without any supervisor or teacher (Chao, 2011). It is normally used for clustering a sample population in different categories with the purpose of segmenting for a particular intervention (Braaten, 2010; Chang et al., 1996; Cutler et al., 2011; Krishnapuram et al., 2005; Liaw & Wiener, 2002; Navot, 2006). Examples of such algorithms include Apriori algorithm and K-means clustering algorithm.
- **Reinforcement Learning:** As its names suggest, it is used to fortify, reinforce or better the function in order to generate the desired results (Bradtke & Duff, 1995 and Chao, 2011). Thus, the machine (computer) is trained to make specific decisions. It is similar to supervised learning but this one lacks certain information about the dataset (Alpaydm, 2010). The machine normally trains itself repeatedly with little help using trial and error mechanism Shalev-Shwartz and Ben-david (2014) and typical example is the Markov Decision Process (Bradtke & Duff, 1995).

2.2.3 Topmost Machine Learning Algorithms

There are so many Machine Learning Algorithms which can be applied to so many dataset problems in areas such as health, engineering , mathematics, business etc. but Le (2018) listed the following as the most topmost and most commonly used Machine Learning

Algorithms:

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes

6. KNN
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms
10. Gradient Boosting algorithms (GBM XGBoost, LightGBM, CatBoost)

Wu et al. (2008) presented a list of top 10 data mining algorithms identified by the IEEE International Conference on Data Mining: C4.5, K-Means, SVM, Apriori, EM, PageRank, AdaBoost, KNN, Naive Bayes, and CART. According to the Wu et al. (2008) these top 10 algorithms are amongst the best powerful data mining algorithms in the scientific research environment. These 10 algorithms cover classification, clustering, statistical learning, association analysis, and link mining, which are all among the most important topics in data mining research and development (Wu et al., 2008).

For the purpose of this study, the focus would be on Decision Tree (C5.0), Random Forest (Breiman and Cutler algorithm) and Neural Network (Back propagation algorithm) which are among the topmost algorithms used in Data Science.

2.2.3.1 Logistic Regression

Logistic regression is one of the supervised Machine Learning classification algorithm used to assign observations to a discrete set of classes (Anouze, 2019; Lim & Yu-Shan, 2000; Ream & Rumberger, 2008 and Wickramasinghe & Karunasekara, 2016). Examples of such classification problems are Email spam or not spam online transactions Fraud or not Fraud, Tumor Malignant or Benign. Logistic regression converts its output using the logistic sigmoid function (Sigmoid function) illustrated in the figure below to return a probability value.

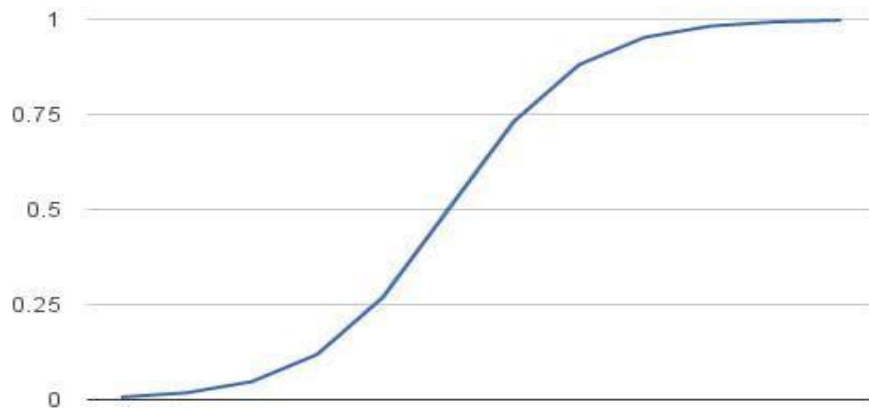


Figure 2.1: A Graph of a Logistic Regression Sigmoid Function (Wickramasinghe & Karunasekara, 2016)

It is a predictive analysis algorithm and based on the concept of probability (Mousavi et al., 2019). The Sigmoid function is an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

$$1 / (1 + e^{-\text{value}})$$

Where e is the base of the natural logarithms (Euler's number or the EXP() function in your spreadsheet) and value is the actual numerical value that you want to transform.

Types of Logistic Regression

Usually, logistic regression means binary logistic regression having binary target variables, but there can be a scenario where two or more groupings of target variables that can be predicted by it. Based on those numbers of groupings, Logistic regression can be divided into following;

- **Binary or Binomial:** With this type of classification, a dependent variable will have only two possible types either 1 or 0. For example, these variables may represent success or failure, yes or no, win or loss etc.

- **Multinomial:** For multinomial classification, dependent variable can have 3 or more possible unordered types or the types having no quantitative significance. For example, these variables may represent “Type A” or “Type B” or “Type C”.
- **Ordinal:** With this type of classification, dependent variable can have 3 or more possible ordered types or the types having a quantitative significance. For example, these variables may represent “poor” or “good”, “very good”, “Excellent” and each category can have the scores like 0,1,2,3.

Representation Used for Logistic Regression

Logistic regression uses an equation as the illustration similar to linear regression. Input values (x) are combined linearly using weights or coefficient values (Beta) to predict an output value (y) (Zhiyu, 2016). A key difference from linear regression is that the output value being modeled is a binary value (0 or 1) rather than a numeric value. Below is an example logistic regression equation:

$$y = e^{(b_0 + b_1 * x)} / (1 + e^{(b_0 + b_1 * x)}) \quad (1)$$

Where y is the predicted output, b₀ is the bias or intercept term and b₁ is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data. The actual representation of the model that you would store in memory or in a file is the coefficients in the equation (the beta value or b's).

Learning the Logistic Regression Model

The coefficients (Beta values b) of the logistic regression algorithm must be protected from your training data. This is done using Maximum-likelihood estimation. The Maximumlikelihood estimation is a common learning algorithm that makes assumptions about the distribution of your data (Kaitlin et al., 2018). The best coefficients would result in a model that would predict a value very close to 1 (e.g. female) for the default class and a value very

close to 0 (e.g. male) for the other class. The intuition for maximum likelihood for logistic regression is that a search procedure seeks values for the coefficients (Beta values) that minimize the error in the probabilities predicted by the model to those in the data (e.g. probability of 1 if the data is the primary class).

2.2.3.2 Decision Tree

A decision tree is a graphical framework representing decisions and their potential outcomes as shown in the Figure 2.1 (Shalev-Shwartz & Ben-david, 2014). It basically consists of three categories of nodes as discussed below.

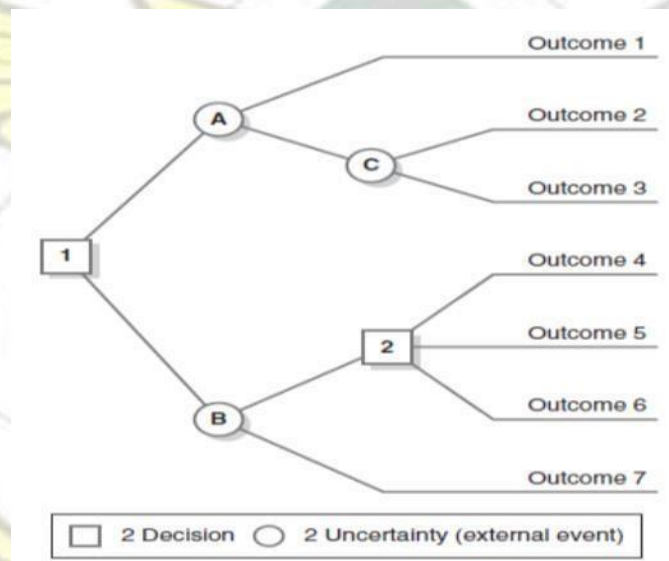


Figure 2.2: A Classic Example of Decision Tree (Shalev-Shwartz & Ben-david, 2014)

1. **Decision node:** These are normally denoted by squares describing decisions that can be suggested. Lines springing from a square indicate all different preferences offered at a node (Jansson, 2016).
2. **Chance node:** This also is symbolized by circles displaying chance outcomes. Chance outcomes are events that can happen but are outside the capability of the decision maker to control (Alpaydin, 2010).

3. **Terminal node:** This last type of node is represented by triangles or by lines having no further decision nodes or chance nodes. Terminal nodes depict the final outcomes of the decision making process (Steinke & Etten, 2017).

Decision Trees are commonly used for classification and regression tree analysis. Decision Trees are becoming increasingly more popular for data mining because they are easy to understand and interpret, require little data preparation, handle numerical and categorical data, and they perform very well with a large data set in a short time (Tso & Yau, 2007) . Decision trees produce excellent visualizations of results and their relationships. Although there are many specific Decision Tree Algorithms, the ID3, C4.5, C5.0, C&RT, and CHAID and QUEST algorithms are the most commonly used ones (Delen et al., 2013). For this study the C5.0 DT algorithm proposed by Ross (1993) was adopted as it offers numerous improvements on C4.5. This C5.0 Decision Tree algorithm has the following features which make it different from other DT algorithms like the ID3, CART and even its immediate successor C4.5 (Delen et al., 2013).

- It normally suggests a dualistic (binary) tree or multi-branches tree
- It also uses Information Gain Entropy, as its splitting principles.
- C5.0 pruning method employs the Binomial Confidence Limit method.
- In an instance of treating omitted values, the C5.0 algorithm agrees whether to approximate the Not Available (NA) values as a function of other attributes or allocates the case statistically among the results.
- The tree produces C5.0 algorithms and its rulesets are typically minor than the tree that C4.5 will produce.
- In terms of **Boosting** while the stochastic gradient boosting machines deviates from the exiting **adaboost** algorithm, C5.0 will do approximately analogous to **adaboost**. After generating the first tree, weights are calculated and consequent iterations create weighted trees or rulesets. The Subsequent trees (or rulesets) are controlled to be about the same size

as the initial model. The final prediction is a simple average of class probabilities created from each tree or ruleset (i.e. no stage weights).

- For **Winnowing**- A feature choosing step done before the model is built. The data set is divided (split) randomly into half and an initial model is fit. Each predictor is removed in turn and the effect on model performance is determined (using the other half of the random split). Predictors are flagged if their removal does not increase the error rate. The final model is fit to all of the training set samples using only the unflagged predictors. R has a library that contains caret function train that has attachments to C5.0 which can tune over the model type, winnowing and even its boosting

Other Features of C5.0

- It is allowed to vary the confidence factor for pruning and also possible to attune the least number of cases in a terminal node
- Turning on possible global pruning algorithm is also permitted □ C5.0 can avoid boosting if it's considered to be unproductive.
- It is also possible to assign unequal costs to precise errors types
- It is more memory efficient faster than its predecessor, C4.5
- It produces analogous outcomes to that of C4.5 with significantly lesser DT
- It's possible to weight distinct cases and misclassification types. **Boosting in C5.0**

As a method to reduce errors in the predictions made by the Decision Trees, a boosting algorithm can be implemented (Jansson, 2016). According to Jansson (2016), the fundamental idea of boosting algorithms is to achieve the following:

- Choose and implement a classifier.
- Split the dataset into a training set (of size n) and a validation set (of size m). The samples in these sets are chosen randomly in the first iteration. Note that both sets need to contain

samples from all classes.

- Train the classifier using the training set. This classifier is said to be a weak classifier. Test the classifier by classifying the testing set and create weights to indicate the flaws in the weak classifier.
- Repeat steps 2 and 3, for a given number of iterations, storing each weak classifier constructed this way. Note that at each iteration, the selection of samples is reflected by the assigned weights.
- Combine all weak classifiers into one strong classifier.

Algorithm 2.2 AdaBoost for binary classification

```

1.  procedure ADABOOST
2.   $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n) \leftarrow n$  observations, where  $\mathbf{x}_j$ , is a vector of attributes
    and  $\mathbf{y}_j$ , is the class.
3.   $\mathbf{w}_t(i) \leftarrow$  weighting function. Initialized as  $1/n$  for  $i = 1, \dots, n$ .
4.   $T \leftarrow$  number of trials
5.  for  $t \leftarrow 1$  to  $T$  do
6.   $f_t(\mathbf{x}_i) \leftarrow$  weak classifier

7.   $\epsilon_t = \sum_{\mathbf{y}_i \neq f_t(\mathbf{x}_i)} \mathbf{w}_t(i)$ 
8.   $\alpha_t = \frac{1}{2} \log\left(\frac{1 - \epsilon_t}{\epsilon_t}\right)$ 
9.  for  $i \leftarrow 1$  to  $n$  do
10.  $\mathbf{w}_{t+1}(i) = \mathbf{w}_t e^{-\alpha_t} \leftarrow$  if correctly classified
11.  $\mathbf{w}_{t+1}(i) = \mathbf{w}_t e^{\alpha_t} \leftarrow$  if incorrectly classified
12. end

13.  $\mathbf{w}_t(i) = \frac{e^{-\sum_{l=1}^t \alpha_l f_l(\mathbf{x}_i)}}{\sum_{i=1}^n e^{-\sum_{l=1}^t \alpha_l f_l(\mathbf{x}_i)}} \leftarrow$  normalize  $\mathbf{w}_t$ 
14. end
15.  $F(\mathbf{x}) = \text{sign}\left(\sum_{t=1}^T \alpha_t f_t(\mathbf{x})\right) \leftarrow$  the final classifier
16. end procedure

```

In C5.0 the particular boosting algorithm implemented is based on the idea of Adaptive Boosting or AdaBoost for short. This work made great impact in the field of machine learning due to the ability of combining several weak classifiers into a stronger one, while retaining the robustness to over-fitting of the weak classifiers.

The main idea of adaptive boosting is to weigh the data points in each successive boosting iteration during the construction of a classifier. These weights for each sample in the training data are distributed such that the algorithm would be focused to correctly classify the data points which were misclassified by the previous classifiers. The differences between the algorithm used in C5.0 and AdaBoost are as follows according to (Jansson, 2016):

- C5.0 tries to maintain a tree size similar to the initial one (which is generated without boosting taken into account). This is correlated with the amount of terminal nodes, which increase in number as the tree grows.
- C5.0 calculates class probabilities for all boosted models and within these models, weighted averages are calculated. Then, from these models, C5.0 chooses the class which has the maximum probability within the group.
- The boosting procedure ends if $\sum_{\{i: y_i \neq F_t(x_i)\}} \omega_t(i) < 0.1$ OR

$$\frac{\sum_{\{i: y_i \neq F_t(x_i)\}} \omega_t(i)}{|W_m|} > 0.5 \quad (1)$$

Where W_m is the cardinality of the set of weights associated with misclassified observations.

The C5.0 weighting algorithm

Algorithm 2.3: The weighting procedure in C5.0 Algorithm

- | | |
|----|--|
| 1. | procedure |
| 2. | $N \leftarrow$ Number of samples in training set. |
| 3. | $N_- \leftarrow$ Number of misclassified samples |
| 4. | $T \leftarrow$ number of boosting iteration. |
| 5. | $w_{i,t} \leftarrow$ Weight of the i-th sample during t-th round boosting. |
| 6. | $S_+ \leftarrow$ Sum over all weights associated with correctly classified samples |

7. $S_- \leftarrow$ Sum over all weights associated with correctly misclassified samples

8. **for** $t \leftarrow 1$ to T **do**

9. ***Build a decision tree***

10. **for** $t \leftarrow 1$ to N **do**

11. $\text{midpoint} \leftarrow \frac{1}{2} [\frac{1}{2} (S_+ + S_-) - S_-]$

12. $W_{i,t} = \frac{W_{i,t-1} S_+ - \text{midpoint} S_+}{S_+ - \text{midpoint}} \leftarrow$ **weight if correctly classified**

13. $W_{i,t} = \frac{W_{i,t-1} S_+ - \text{midpoint} S_-}{S_+ - \text{midpoint}} \leftarrow$ **weight if misclassified**

14. **end** 15. **end** 16. **end procedure**

The Decision Tree is one of the topmost machine learning algorithms which suggests a graphical or diagrammatical illustration of a technique for classifying, predicting and evaluating an item of importance or concern (Jain et al., 2016a). It is an easy and commonly used classification method. It deals with decision analysis by employing a tree-like structure of decisions and their relative potential outcomes (Hu & Wang, Shuaiwei 2016). It has nodes where at each node in a Decision Tree an attribute must be selected to divide the node's instances into subgroups. They are also a type of supervised learning method which splits dataset into more standardized clusters as possible from the variable to be predicted. Decision Tree accepts input set of well-ordered data, and output shaft is delivered in which each end node (leaf) is a decision (a class) and each non-end node (middle) shows a test (Hssina et al., 2016). They are normally used for acquiring facts to aid in decision making. It normally begins with a root node for the users to take the necessary actions then from the node the user given individual node recursively based on an adopted DT algorithm. The end product is a tree with each branch denoting a potential scenario of the decision and its conclusion (Ogunde & Ajibade, 2014).

The most commonly used algorithms in DT are: ID3, CART, CHAID and C4.5 with its extension C5.0. For this study, the C4.5 Algorithm which is an extension of the ID3 proposed by Ross Quinlan in 1994 (Pandya & Pandya, 2015) was adopted and implemented in R studio using R codes with package “RWeka” (Hornik et al., 2019).

2.2.3.3 Random Forest

Random Forests (RF), on the other hand, are an algorithm for classification proposed by Breiman in 2001 and cited by Chi et al.(2011) and Cutler et al.(2011) that utilize an ensemble of classification trees. Random Forest is a simple, not complex to use Machine Learning Algorithm that creates, even without hyper-parameter tuning, a good result (Cutler, 2010). It happens to be one of the topmost and most used algorithms, because it is simple and can also be applied in both classification and regression tasks (Le, 2018). It is also one of the supervised Machine Learning Algorithm. As the name implies, it produces a forest and makes it random, “Forest”. This makes it a combination of trees and most at time trained with the “bagging” method (Breiman, 2001a; Chao, 2011; Liaw & Wiener, 2002; Scornet, 2010). The main idea behind this bagging method is that, a combination of learning models will increase the overall result (Breiman, 2001a; Chao, 2011) as shown in the Figure 2.2.

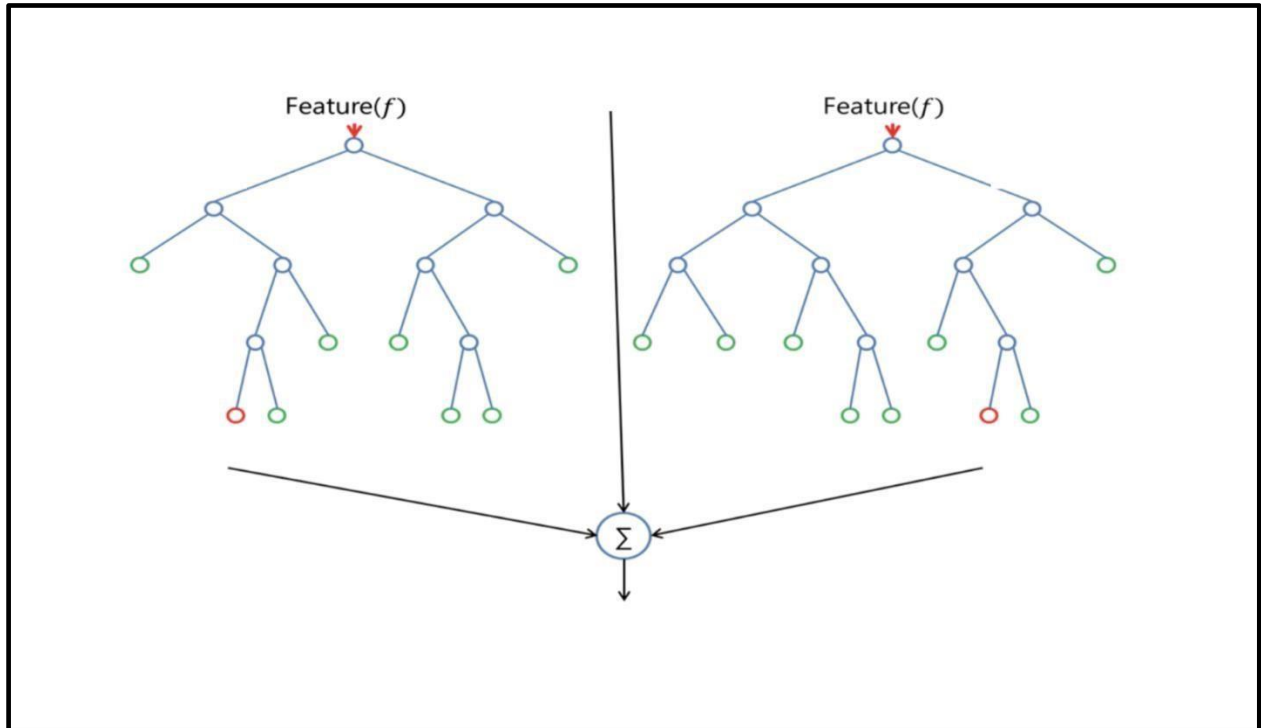


Figure 2.3: A Classic Example of Random Forest (Cutler et al., 2011)

Random Forests (RF) are algorithm for classification developed by Breiman in 2001 and cited by Chi et al. (2011) that utilize an ensemble of classification trees (Cutler et al., 2011; Cutler, 2010 and Gislason et al. (2006). RF is an ensemble machine learning algorithm (Cutler et al., 2011).

The fundamental principle of the RF algorithm is that constructing a smaller DT with limited characteristics is inexpensive process in terms of computation (Cutler et al., 2011). Thus, it is possible to construct numerous small, weak Decision Trees in parallel and merge these smaller trees to form one strong learner by using their mean performance or even or selecting the popular one. In terms of application and practicability, RF algorithms are considered to be most precise learning algorithms to date (Cutler et al., 2011).

The RF algorithm adopted for this study was Leo Breiman and Adele Cutler Random Forest Algorithm Breiman (2001) and Cutler et al. (2011) illustrated in algorithm 2.7. This was implemented in R using R codes with “randomForest” package (Liaw & Matthew, 2018 and

Tang, 2016). According to literature Random Forest has achieved a lot of good classifications and other prediction. A Random Forest model has a better ability in modeling and predicting. An important feature of Breiman's algorithm according to Livingston (2005) is the variable importance calculation. The success of the RF model is attributed to its generality ability to forecast the output for new data after the RF has been trained with a similar dataset. Literature also shows that RF as a Machine Learning Algorithm delivers higher classification and predictive accuracies than statistical procedures. Finally random forest algorithm utilizes the bagging method for building an ensemble of Decision Trees. Bagging is known to reduce the variance of the algorithm.

Random Forest can be used for any variable being it categorical target variable normally considered as "classification", or a continuous target variable, also called "regression". The same applies to the independent or predictor variables (Cutler et al., 2011).

Random Forest is a tree-based ensemble where each individual tree is subject to an assembly of random variables (Cutler et al., 2011). Thus, for instance, using a p -dimensional random vector $X = (X_1, \dots, X_p)^T$ indicating the independent or predictor variables and a random variable Y denoting target or dependent or response variable, one can suggest an unidentified combined distribution $P_{XY}(X, Y)$. The purpose is to select a forecast function $f(X)$ for predicting Y . The prediction function is assessed by a loss function $L(Y, f(X))$ and formulated to reduce the anticipated value of the loss

$$E_{XY}(L(Y, f(X))) \quad (2)$$

where the XY as subscripts represent expectancy with deference to the combined distribution of X and Y .

Instinctively, $L(Y, f(X))$ is a quantity of how near $f(X)$ is to Y ; it disciplines values of $f(X)$ that are far from Y . Archetypal selections of L are *squared error loss* $L(Y, f(X)) = (Y - f(X))^2$ for regression and *zero-one loss* for classification:

$$L(Y, f(x)) = I(Y \neq f(x)) = \begin{cases} 0 & \text{if } Y = f(x) \\ 1 & \text{if } Y \neq f(x) \end{cases} \quad (3)$$

1 otherwise

It happens that reducing $E_{XY}(L(Y, f(X)))$ for squared error loss gives the provisional anticipation

$$f(x) = E(Y|X = x) \quad (4)$$

Else recognized as the regression function. In case of the classification, if the list of likely values of Y is designated by \mathcal{Y} , reducing $E_{XY}(L(Y, f(X)))$ for zero-one loss gives

$$f(x) = \arg \max_{y \in \mathcal{Y}} P(Y = y|X = x) \text{ popularly called the Bayes rule.} \quad (5)$$

Ensembles build f using an assembly of “base learners” $h_1(x), \dots, h_J(x)$ and the base learners are joined to provide the “ensemble predictor” $f(x)$. With respect to regression, the mean of the base learners used

$$f(x) = \frac{1}{J} \sum_{j=1}^J h_j(x) \quad (6)$$

Whereas in classification $f(x)$ is the class with the highest predictions (voting)

$$f(x) = \arg \max_{y \in \mathcal{Y}} \sum_{j=1}^J I(y = h_j(x)) \quad (7)$$

The j th base learner j th in RF is a tree designated $h_j(X, \odot_j)$, where \odot_j is a assembly of random variables and the \odot_j 's are independent for $j = 1, \dots, J$.

2.2.3.4 Artificial Neural Network

Artificial Neural Network popular called ANN or Neural Network (NN) is a computational framework built on the architecture and also operates like the human biological neural networks (Lam, 2004). The flow of information through the network impact on the structure of the ANN because a neural network changes or learns, based on the input and output (Emrouznejad & Shale, 2009; Gal, 2016; Ko, 2004; Krishnapuram et al., 2005). They are classified as nonlinear statistical data modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found (Cutler et al., 2011 and Ngai et al., 2009). Artificial Neural Network (ANN) is a type of artificial intelligence technique that mimics the behavior of the

human brain (Lam, 2004). A Neural Network is a massively parallel distributed processor made up of simple processing units that have a natural tendency for storing experiential knowledge and making it available for us. ANNs can be grouped into two major categories: feed-forward and feedback (recurrent) networks. In the former network, no loops are formed by the network connections, while one or more loops may exist in the latter. The most commonly used family of feed-forward networks is a layered network in which neurons are organized into layers with connections strictly in one direction from one layer to another (Shamisi et al., 2011). The basic system of NN without the hidden layer consists of only two layers, the input and output layer. This normally called the skip layer because it is made up of a straightforward linear regression modeling in a ANN design. The input layer communicates directly with the output layer without involving the hidden layer. The figure 3.6 below shows a classical one layer NN with a hidden layer (Anantwar & Shelke, 2012).

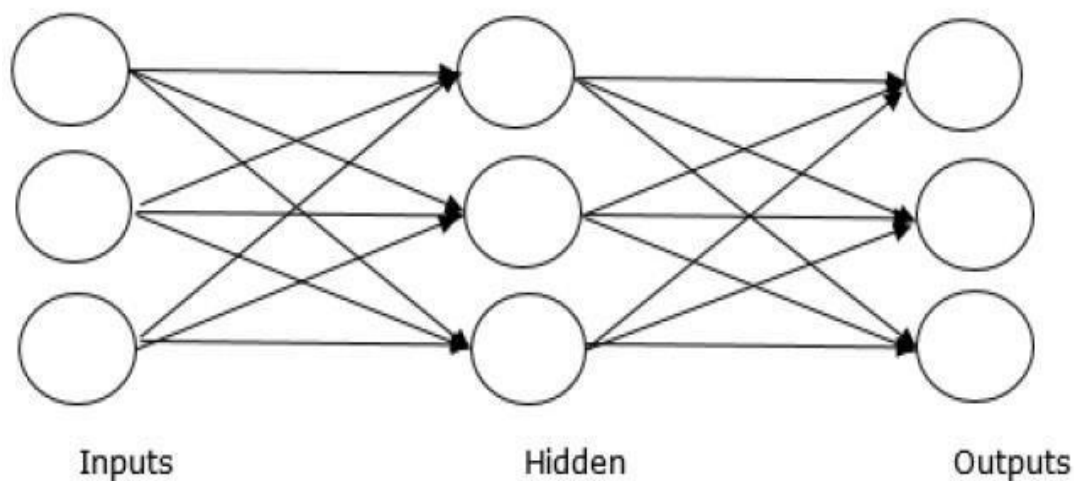


Figure 2.4: Classical NN model (Shamisi et al., 2011)

Multilayer Layer Perceptron (MLP)

Neural Networks consist of a large class of different architectures. In many cases, the issue is approximating a static nonlinear, mapping $f(x)$ with a neural network $f_{NN}(x)$, where $x \in \mathbb{R}^K$ (Koivo, 2008). The most common and important ANN in function approximations are Multilayer Layer Perceptron (MLP) and Radial Basis Function (RBF) networks (Koivo, 2008; Mostafa, 2009; Shamisi et al., 2011). The MLP network is employed in the present study.

The MLP comprises an input layer with two or more hidden layers, and an output layer. Node i , also called a neuron, in a MLP network is shown in Figure 3.7 below. This also comprises of a summation(Σ) and a nonlinear activation function g . It is important to note that NN where the hidden neurons have sigmoidal activation function and the output neurons the sigmoidal or identity function are called Multi-Layer Perceptron (Shamisi et al., 2011).

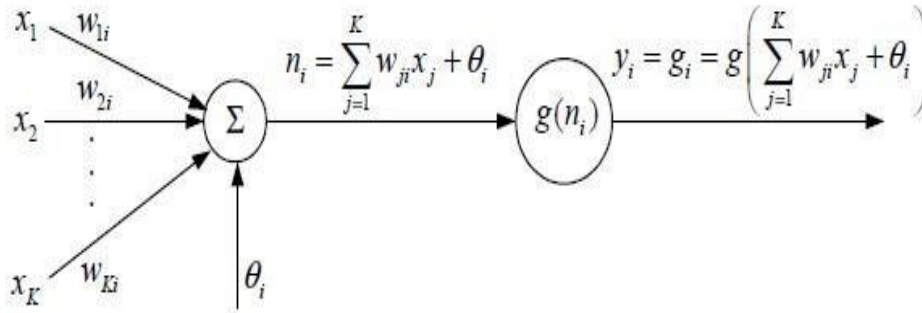


Figure 2.5 : Single node in a MLP network source (Koivo, 2008)

The inputs $x_k, k = 1, \dots, K$ to the neuron are multiplied by weights w_{ki} and summed up together with the constant bias term θ_i . The resulting n_i is the input to the activation function g . The activation function was initially selected to be a transmit function, but mathematical expediency a hyperbolic tangent (tanh) or a sigmoid function are most commonly used. The Hyperbolic tangent is formulated as:

$$\tanh(x) = \frac{1 - e^{-x}}{1 + e^x} \quad (8)$$

The output of node i is given by,

$$y_i = g_i = g\left(\sum_{j=1}^K w_{ji} x_j + \theta_i\right) \quad (9)$$

Linking two or more nodes in parallel and also in series gives a MLP network with an example shown in the figure below. The same activation function g is used in both layers. The superscript of n, θ or w refers to the layer, first or second. The output $y_i, i = 1, 2$, of the MLP network becomes

$$y_i = g\left(\sum_{j=1}^3 w_{ji2} g(n_{j1}) + \theta_{j2}\right) = g\left(\sum_{j=1}^3 w_{ji2} g\left(\sum_{k=1}^K w_{kj1} x_k + \theta_{j1}\right) + \theta_{j2}\right) \quad (10)$$

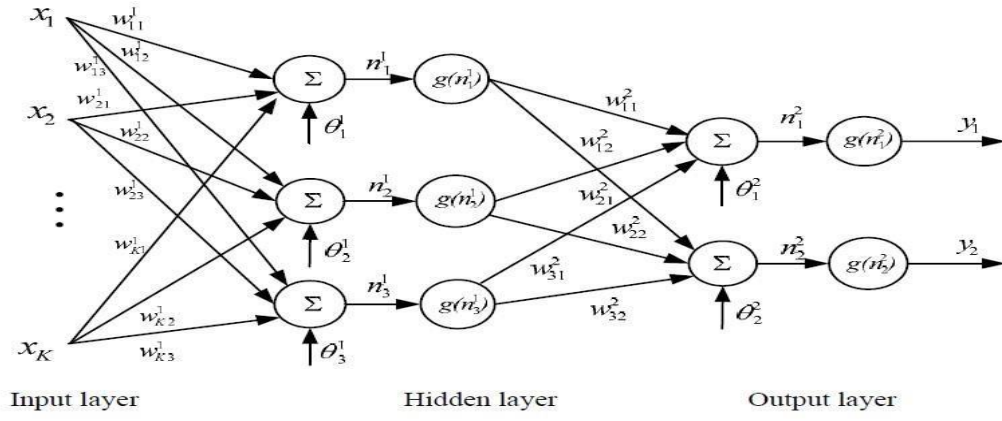


Figure 2.6: A MLP network with one hidden layer source (Koivo, 2008)

From (3) we can conclude that a MLP network is a nonlinear parameterized map from input space $\mathbf{x} \in \mathbf{R}^K$ to output space $\mathbf{y} \in \mathbf{R}^m$ (where $m = 3$ in this case). The parameters are the weights w_{ji}^k and the biases θ_j^k . Activation function g is normally presumed to be equal in all layer and acknowledged in advance. In the figure the same activation function g is used in all layers. Given dataset $(x_i, y_i), i = 1, \dots, N$, calculating the best MLP network is expressed as a data fitting problem. The parameters to be determined are (w_{ji}^k, θ_j^k) . Building and processing of the ANN depends on the following building blocks.

Fixing the structure of MLP Network Topology: the number of hidden layers and neurons (nodes) in each layer.

Activation Functions: The activation functions for each layer are also chosen at this stage, that is, they are assumed to be known.

The unknown parameters to be estimated are the weights and biases, (w_{ji}^k, θ_j^k) .

There are many existing algorithms for determining these network parameters but according

to literature on Neural Network the algorithms are called learning or teaching algorithms

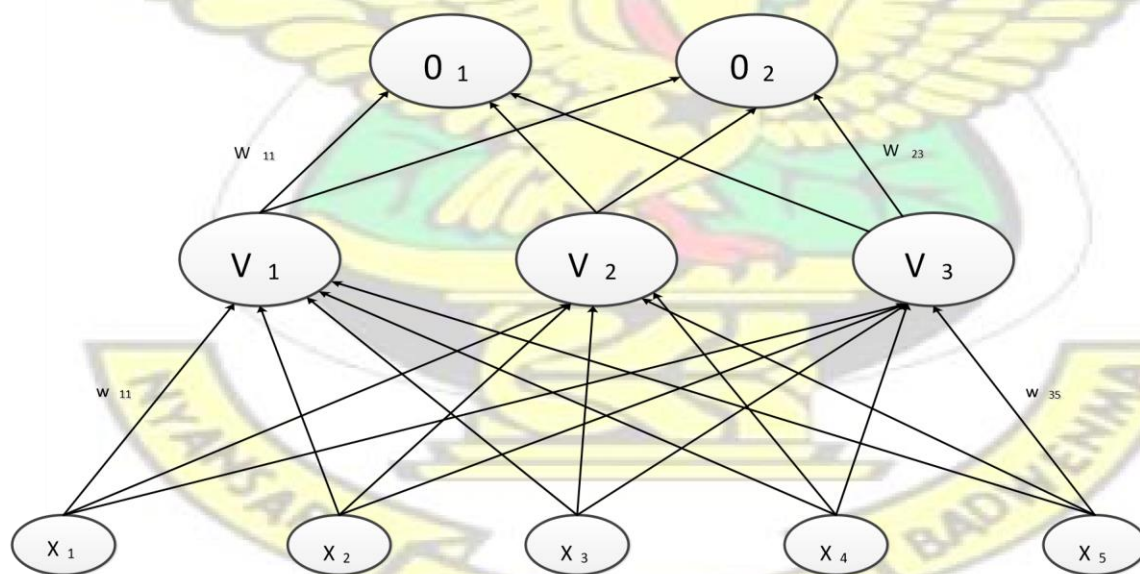
(Tsai et al., 2009). In system identification, they belong to parameter estimation algorithms (West, 2000). The most well-known are back-propagation and Levenberg-Marquardt

Algorithms (Koivo, 2008).

Back-propagation is a gradient based algorithm, which has many variants (Shamisi et al., 2011). Levenberg-Marquardt is usually more efficient but needs more computer memory (Koivo, 2008).

2.2.3.5 The Back-Propagation Algorithm

A back propagation network is the most popular network because of its capability to find non-linear solutions to undefined complex problems (Kriesel, 2005). The errors of the output of a back propagation network are propagated to back by means of the same connections used in feed forward mechanism by the derivation of the feed forward transfer function (Hinton, 2018). The learning function in this network is based on a basic two-way memory incorporation (Ekici & Aksoy, 2009). The algorithm gives a prescription for changing the weights w_{ij} in any feedforward network to learn a training set of input output pairs $\{x_d, t_d\}$.



Let us consider a simple two-layer MLP network shown Figure 2.5.

O_i

w_{ij}

V_j

w_{jk}

X_k

Figure 2.7: A two-layer MPL Network

Given the pattern X_d the hidden unit j receives a net input,

$$net_j^d = \sum_{k=1}^5 w_{jk} x_k^d \quad (11)$$

$$\text{And gives out the output, } V_j^d = f(net_j^d) = f(\sum_{k=1}^5 w_{jk} x_k^d). \quad (12)$$

The output unit in this case receives

$$net_i^d = \sum_{j=1}^3 W_{ij} V_j^d = \sum_{j=1}^3 (W_{ij} \cdot f(\sum_{k=1}^5 w_{jk} x_k^d)) \quad (13)$$

and generate the final output

$$O_i^d = f(net_i^d) = f(\sum_{j=1}^3 W_{ij} V_j^d) = f(\sum_{j=1}^3 (W_{ij} \cdot f(\sum_{k=1}^5 w_{jk} x_k^d))) \quad (14)$$

$$\text{The out usual error function is } E[W^{\leftrightarrow}] \equiv \frac{1}{2} \sum_{d \in D} (t_d - O_d)^2 \quad (15)$$

For i outputs and m inputs pairs $\{X_d, t_d\}$

$$E[W^{\leftrightarrow}] = \frac{1}{2} \sum_{d=1}^m \sum_{i=1}^l (t_i^d - O_i^d)^2 \quad (16)$$

Applying the example in this case, E will be formulated as

$$E[W^{\leftrightarrow}] = \frac{1}{2} \sum_{d=1}^m \sum_{i=1}^2 (t_i^d - O_i^d)^2 \quad (17)$$

$$E[W^{\leftrightarrow}] = \frac{1}{2} \sum_{d=1}^m \sum_{i=1}^2 (t_i^d - f(\sum_j W_{ij} \cdot f(\sum_{k=1}^5 w_{jk} x_k^d)))^2 \quad (18)$$

If $E[W^{\leftrightarrow}]$ is differentiable given f is differentiable then gradient descent can be applied

For hidden-to-out connections the gradient decent rule gives:

$$\Delta W_{ij} = -\eta \frac{\partial E}{\partial W_{ij}} = -\eta \sum_{d=1}^m (t_i^d - O_i^d) f'(net_i^d) \cdot (-V_j^d) \quad (19)$$

$$\Delta W_{ij} = \eta \sum_{d=1}^m (t_i^d - O_i^d) f'(net_i^d) \cdot V_j^d \quad (20) \quad \delta_i^d$$

$$= f'(net_i^d) (t_i^d - O_i^d) \quad (21)$$

$$\Delta W_{ij} = \eta \sum_{d=1}^m \delta_i^d V_j^d \quad (22)$$

We can use the chain rule to differentiate with respect to W_{ij} to obtain the input-to hidden connection W_{ij}

$$\Delta W_{jk} = -\eta \frac{\partial E}{\partial W_{jk}} = -\eta \sum_{d=1}^m \frac{\partial E_{Vjd}}{\partial W_{jk}^d} \cdot \frac{\partial W_{jk}^d}{\partial W_{jk}} \quad (23)$$

$$\Delta W_{jk} = \eta \sum_{d=1}^m \sum_{i=1}^2 (t_{id} - o_{id}) f'(net_{id}) \cdot W_{ij} f'(net_{id}) \cdot x_{kd} \quad (24)$$

$$\delta_i^d = f'(net_i^d) (t_i^d - o_i^d) \quad (25)$$

$$\Delta W_{jk} = \eta \sum_{d=1}^m \sum_{i=1}^2 \delta_{id} \cdot W_{ij} f'(net_{id}) \cdot x_{kd} \quad (26)$$

$$\delta_{jd} = f'(net_{jd}) \sum_{i=1}^2 W_{ij} \delta_{id} \quad (27)$$

$$\Delta W_{jk} = \eta \sum_{d=1}^m \delta_{jd} \cdot x_{kd} \quad (28)$$

$$\Delta W_{ij} = \eta \sum_{d=1}^m \delta_{jd} V_j^d \quad (29)$$

$$\Delta W_{jk} = \eta \sum_{d=1}^m \delta_j^d \cdot x_{kd} \quad (30)$$

This will result in the same form with different definitions of δ

Generally, with an arbitrary number of layers, the back-propagation update rule has always the formulae $\Delta W_{ij} = \eta \sum_{d=1}^m \delta_{output} \cdot V_{input}$ (31) Where output and input refers to the connection under consideration. V stands for the appropriate input (hidden unit or real input, x_d) and δ depends on the layer under consideration.

Equation (41) will help us to calculate for a given hidden unit V_j in terms of the δ 's of the unit O_i . The coefficient are usual forward, but the errors δ are propagated backward as shown in the figure 2.6 below using dotted lines.

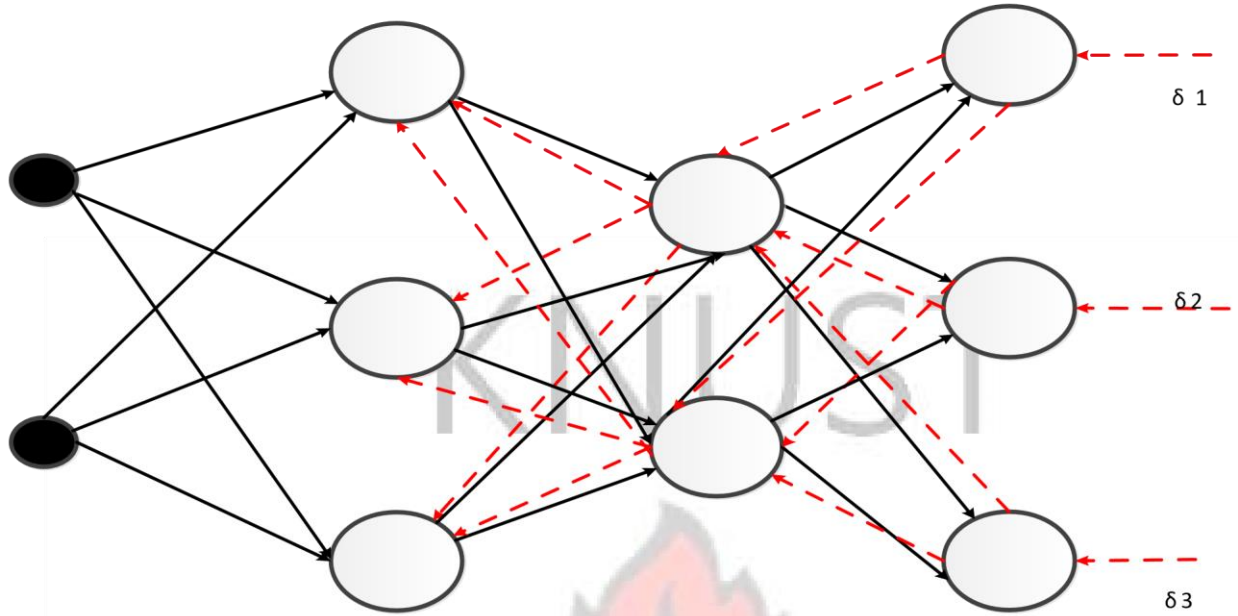


Figure 2.8 NN model with back -propagated errors

We have to use a nonlinear differentiable activation function such as

$$f(x) = \sigma(x) = \frac{1}{1+e^{(-\alpha x)}} \quad (32)$$

$$f'(x) = \sigma'(x) = \alpha \cdot \sigma(x) \cdot (1 - \sigma(x)) = \tan(\alpha \cdot x) \quad (33)$$

$$f(x) = \alpha \cdot (1 - f(x)^2) \quad (34)$$

2.2.4 Evaluation of Machine Learning Algorithms Models

The evaluation of any Machine Learning Algorithm being it supervised, unsupervised or a reinforcement is made a critical aspect of any computer experiment (Kotsiantis, 2007). For instance, if a particular model produces a substantial outcome when evaluated using an accuracy-score may not necessary means that the same model when evaluated against other metrics logarithmic-loss may also give a good result. This means that the selection of metrics influences how the performance of machine learning algorithms is measured and compared. They influence how one sees the importance of different characteristics in the results and your ultimate choice of which algorithm to choose (Kuncheva & Whitaker, 2003).

In most literatures, authors normally use classification accuracy to assess model performance.

Nevertheless, this is not enough to really assess the performance of a model. In this section, the thesis discussed the various metrics that can be used to assess or evaluate the performance of a machine learning algorithm or model. There are so many metrics for evaluating machine learning algorithms but for the purpose of this review, the thesis focuses on the following:

Classification Accuracy: Classification Accuracy is actually the meaning of the term

“Accuracy” in machine learning performance measure (Osisanwo et al., 2017).

Mathematically, it is defined as the ratio of the number of perditions done correctly by the machine learning algorithm to the total dataset that was used as inputs.

$$\text{Accuracy} = \frac{\text{Number of correct Predictions}}{\text{The Total number of predictions made}} \quad (35)$$

This would normally work accurately whenever there are equal numbers of sample belonging to each class. For instance, if there are 98% observations in a particular class, say class 1 and 2% observations in class 2 in our training set, then, it is most likely that the proposed model can easily predict at an accuracy of 98% .This is just a matter of predicting all observations in the training dataset belonging to class 1. In another scenario, if the same machine learning model is applied on a dataset with 70% observations of class1 and 30% observations of class 2, then the accuracy would most likely reduce to 60%. Classification Accuracy is a good measure but at times it generates poor sense of realizing high accuracy (Osisanwo et al., 2017). The reality would be much expensive when the cost of misclassification of the minor (in our case class 2) is very high.

- Logarithmic Loss: Logarithmic Loss or Log Loss also operates by disciplining the false classifications (Kubat et al., 1998). It performs better for multi-class classification (Kuncheva & Whitaker, 2003). With this approach the classifier would allocate probability to each class for all the samples. Assuming, there are N samples in each M classes then the Log Loss is calculated as below:

$$\text{Log Loss} = -\sum_{i=1}^N \sum_{j=1}^M y_{ij} * \log(p_{ij}) \quad (36)$$

Where

y_{ij} , indicates whether sample i belongs to class j or not

p_{ij} indicates the probability of sample i belonging to class j

Log Loss has no upper bound and it exists on the range $[0, \infty]$. Log Loss closer to 0 indicates higher accuracy, whereas if the Log Loss is not close to 0 then it shows poor accuracy.

General, reducing Log Loss offers better accuracy for the classifier.

- **Confusion Matrix:** Confusion Matrix as also presents a matrix as output and defines the comprehensive performance of the model. There are 4 important terms:

True Positives: The scenario where the model predicted YES and the real response was also YES.

True Negatives The scenario where the model predicted NO and the real response was YES.

False Positives: The scenario where the model predicted YES and the real response was NO.

False Negatives: The scenario where the model predicted NO and the real response was also NO.

In this case the accuracy for the matrix can be evaluated by finding the mean of the values lying across the “**main diagonal**” i.e.

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{FalseNegatives}}{\text{Total Number of Samples}} \quad (37)$$

The confusion Matrix forms the basis for the other types of metrics.

- **Area Under Curve:** The Area Under Curve (AUC) is the most extensively used metrics for evaluating machine learning algorithms (Bradley, 1997). It is applied in binary classification problem. AUC of a classifier is the same as the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative

example (Daisuke & Perez, 2017). The following two basic terms are normally used in AUC:

True Positive Rate (Sensitivity): True Positive Rate is defined as $TP / (FN + TP)$. True Positive Rate represent the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

$$\text{TruePositiveRate} = \frac{\text{TruePositive}}{\text{FalseNegative} + \text{TruePositive}} \quad (38)$$

False Positive Rate (Specificity): False Positive Rate is defined as $FP / (FP + TN)$. False Positive Rate represents the proportion of negative data points that are incorrectly considered as positive, with respect to all negative data points.

$$\text{TruePositiveRate} = \frac{\text{FalsePositive}}{\text{FalsePositive} + \text{TrueNegative}} \quad (39)$$

For instance, Daisuke & Perez (2017) in an attempt to predict future firms performance using Machine learning and dataset from Japan utilize the ROC curve as to evaluate the predictive performance of the model. According to the authors, their tasks of binary exit, growth, and profit growth classification require the setting of thresholds for which predicted probabilities surpassing this level will indicate a positive binary outcome. Given a fixed model, the ROC curve plots the true and false positive rates corresponding to the varying of this threshold value. Without any predictors (i.e. random guess), the curve should trace the 45-degree line, and curves closer to the top-left corner are desirable (maximize true positive rate and minimize false positive rate). With this motivation, it is conventional to also summarize the ROC curve by the area under the curve, called AUC (Daisuke & Perez, 2017).

- **F1 Score:** This is used to evaluate a test's accuracy. Is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It determines how precise the classifier is as well as how robust it is. A higher precision but lower recall, offers an enormous accuracy, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. F1 Score attempts to find the balance between precision and recall.

Mathematically:

$$F^1 = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}} \quad (40)$$

Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositives} \quad (41)$$

Recall: It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). Specificity is the exact opposite of Recall.

Mean Absolute Error: Mean Absolute Error is the mean of the difference between the Original Values and the Predicted Values. It offers the measure of how the predictions deviated from the actual output but do not suggest an information about the direction of the error; Thus whether we are under predicting the data or over predicting the data.

$$MeanAbsoluteError = \frac{1}{N} \sum_{j=1}^N |y_j - \hat{y}_j| \quad (42)$$

- **Mean Squared Error (MSE):** This is almost the same us Mean Absolute Error but the only difference being that MSE takes the mean of the square of the difference between the original values and the predicted values (Patel et al., 2014b). The advantage of MSE is that

it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient (Fallahpour et al., 2017). As, we take square of the error, the effect of larger errors become more pronounced than smaller error, hence the model can now focus more on the larger errors (Tavana et al., 2016).

$$MeanSquaredError = \frac{1}{N} \sum_{j=1}^N (y_j - \hat{y}_j)^2 \quad (43)$$

Yousaf (2016) in a study to assess factors that are likely to predict the employee turnover built a Machine Learning Algorithm to do that task. To evaluate the model's performance, the author used overall accuracy, kappa, sensitivity, specificity and ROC Curve as the model's evaluation metrics. According to the authors these machine learning performance metrics are considered best practices for doing such classification task (Yousaf, 2016).

2.2.5 Application of Machine Learning Algorithms

Álvarez (2016); Prakash et al. (2017); Bani-Hani et al. (2009); Corrado et al. (2014); Ko & Kweku-Muata (2014); Ko & Osei-Bryson (2004) and Koellinger (2005) have since embarked on studies that are based on Machine Learning Algorithms. For example, Ko & Osei-Bryson (2004) used a new method called Multivariate Adaptive Regression Splines (MARS) which they suggested will provide supplementary understandings on the characteristics of the impact of IT investments on productivity. The results of this new study were compared with findings from previous works with the same data set. While the results of the former study suggested a positive but uniform impact of IT on productivity, this new study suggested that IT impact on productivity was not uniform but is dependent on other supporting factors. Ko & Osei-Bryson (2004) concluded that there was a complementary relationship that exists between IT and nonIT related investments and stated that additional investment may not result in higher organizational productivity.

Ko & Kweku-Muata (2014) revisited this issue and reconsidered the impact of IT capital on hospital productivity by using two data mining methods. According to the authors, this new approach gave them the strength to discover interactions between the input parameters as well as provisional impacts. The results suggested a complex relationship between IT investment and productivity. The descriptive and regression analysis of data on Small and Medium Enterprises from the two selected countries in Africa (Kenya and Tanzania) suggested that investment in ICT is one vital element of total factor productivity once a definite threshold is passed (Wolf, 2001). Commander et al. (2011) also used a regression analysis on a data on manufacturing companies in both Brazil and India suggested a strong positive relationship between ICT investment and productivity in both countries. Alinezhad (2016); Lee (2010) and Rocha & Júnior (2010) all in one way or the other have applied a DT to make classifications or predictions regarding firms.

Carmona et al. (2018) predicted bank failures in US using an extreme gradient boosting approach which was based on 156 banks and their financial ratios. The results indicated that smaller values for financial ratios such as retained earnings to average equity, pretax return on assets, and total risk-based capital ratio were linked with a higher risk of bank failure. Chi et al. (2011) in a research proposed a credit rating forecasting model using market-based information as an independent variable. The Moody's KMV (KMV) model was used to assess the market-based information of each business. In order to validate the suggested technique, the authors used a hybrid model, which comprises random forests (RF) and Rough Set Theory (RST) to excerpt valuable data for credit rating. The results demonstrated that market-based data provides significant information in credit rating forecasts. Creamer (2009) utilized random forest and logistic regression for performance forecast of Latin America ADRs and banks. The study also arranges in order of magnitude accounting and corporate variables based on their impact on performance. According to Creamer (2009), Random Forest was intelligent to specify the most significant variables. Creamer (2009) finally concluded that elucidation of

predictive models for smaller dataset enhanced when the ability of Random Forest to rank and predict with the parameters of a logistic regression was integrated. Booth et al. (2014) proposed a special framework that utilizes new Machine Learning methods to forecast the price return over these seasonal events, and then use the predictions to build a profitable trading strategy. Kartasheva & Traskin (2011) utilized an adapted Random Forest classification algorithm to predict insolvency of insurers. The study disclosed that RF methodology provides higher quality of prediction compared to other existing methods. Daisuke & Perez (2017) in another study with the main aim to predict new companies efficiency with Machine Learning procedures used data on a large number of Japanese firms with supply-chain linkage information provided by a credit reporting agency. Using Random Forest Random, the study showed high accuracy in prediction. The authors suggested that the evidence of theory of the study offers real-world usage of Machine Learning methods in firm efficiency forecast.

Bia et al. (2006) discussed a new numerical method to predict growth in manufacturing from firm-survey responses. The authors based their forecasts on a predicting algorithm motivated by the Random Forest which is fast, strong to noise and permits for the handling of missing values. Tanaka et al. (2016) presented an original Random Forests-based early warning system for forecasting bank failures. The results indicated that the RF outpaces ordinary methods regarding forecast precision. Biau et al. 2006; Luca et al. (2010) and Patel et al. (2014) all have contributed to the Random Forest and organizations literature.

According to literature, NN has been utilized in different ways for classification and predictions because it has a high computational power and performance (Gemino & Sauer, 2010 and Lam, 2004). For instance, Boritz & Kennedy (1995) examined the efficiency and performance of a number of neural networks in predicting bankruptcy filing. The authors considered two methods for training Neural Networks namely: Back-Propagation and Optimal Estimation Theory. For the back-propagation training technique, four distinct models namely, BackPropagation, Functional Link Back-Propagation With Sines, Pruned BackPropagation,

and Cumulative Predictive Back-Propagation (Kriesel, 2005) were tested. The authors did a comparative analysis of Neural Networks against the ordinary bankruptcy prediction practices such as Discriminant Analysis, Logit, and Probit.

The granting of loans to customers is one of the main functions of a credit union. Consequently, the application of technological tools that can support such business process is necessary and could be a key element in credit management (Sousa & Figueiredo, 2014). Sousa & Figueiredo (2014) developed two models that comprise Neural Network and Decision Tree models to analyze the capability of a credit union's customers to settle their debts. The results were evaluated through cross-validation of ten sets, repeated in ten simulations. The authors concluded that the models have statistically similar results and may help in a cooperative's decision-making process especially in credit management.

Lam (2004) also investigated the ability of Neural Networks, specifically back propagation algorithm and suggested a high accuracy of NN in predictions.

To improve the accuracy and the computation speed of credit scoring models, (Sang, Nam, & Nhan, 2016) suggested a credit scoring model which was based on parallel Random Forest classifier and feature selection method to evaluate the credit risks of applicants. By integrating Random Forest into feature selection process, the authors indicated that importance features can be accurately evaluated to remove irrelevant and redundant features. The authors proposed an algorithm to select best features by using the best average and median scores and the lowest standard deviation as the rules of feature scoring. They experimentally assessed the performance of their algorithm by using two public datasets from Australia and Germany. The results of their study suggest an improved accuracy of their model comparable to other commonly used feature selection methods particularly in terms of their prediction accuracy, 76.2% with a significantly reduced running time of 72 minutes on German credit dataset and the highest average accuracy of 89.4% with the running time of only 50 minutes on Australian credit dataset (Sang et al., 2016).

According to Roy & Urolagin (2019), the assessment of credit risk within the banking sector has become an important issues. Roy & Urolagin, (2019) therefore, proposed a methodology made up of both Random Forest and Support Vector Machine that performs a two-level data processing in order to accurately predict creditworthiness of the clients. According to the authors, the proposed methodology would aid to attain results with minimized false positives. Despite the countless benefits of Automated Teller Machines (ATMs) to customers within the banking sector, ATM lacks the provisioning of security measures against frauds (Bajaj et al., 2019). Video surveillance has also been suggested to be one of the protuberant measures against ATM frauds (Bajaj et al., 2019). Bajaj et al. (2019) presented a method that can be applied for activity recognition in small premises such as ATM rooms by encoding the motion in images. The authors utilized Gradient-based descriptor (HOG) to extract features from image sequences and classified the extracted features using random forest classifier. Bajaj et al. (2019) concluded that their method was positive in detecting abnormal and normal human activities both in case of single and multiple personnel with an average accuracy of 97%. A combine Random Forests and Logistic Regression was used by (Germán Creamer, 2009) for performance prediction of Latin American ADRS and Banks and suggested that the most important variables that affect banks are size, long term assets to deposit, number of directors and the efficiency of the legal system.

Finally, Emrouznejad & Shale (2009) ; Hsu et al. (2016) ; Ma et al. (2018) ; Nami & Shajari (2018) ; Portas & Abou-Rizk (1997) ; Toloo et al. (2015) ; Tsai et al. (2009) ; Tzeng (2014) ; West (2000) ; Wu (2009) and Yeo & Grant (2018) have all contributed to the Machine Learning Algorithm literature.

2.3 APPLICATION OF COMBINED DEA AND MACHINE LEARNING ALGORITHM

According to literatures, most studies either combine two or more Machine Learning Algorithms or integrate it with other models such as the popular non-parametric DEA model to detect or predict an outcome. For instance, Sakouvogui (2019) applied an integrated DEA with Support Vector Machines (SVM) on three-monthly dataset of US Agricultural Banks and

indicated a strong evidence that the efficiency measures of the selected banks in the dataset were strong before the financial crisis (2005-2006), during the financial crisis (2007-2009) and even after the financial crisis (2010-2016). The author further suggested that the integrated DEA-SVM had a lower performance during 2007-2009. Lee (2010) applied a combined DEA and Decision Tree for efficiency analysis and suggested appropriate guidelines for controls in Business- to-Consumer (B2C) applications. The author indicated that retail firms and information service providers implement B2C controls more efficiently than financial firm and suggested the possibility of using Decision Trees for controls assessment in B2C applications.

According to (Jain et al., 2016b), in most cases, the choice of input and output used for DEA assessment was based on the researchers' own judgments. Which according to them, most researchers lack the rigorous justification relating to the suitability of such input/output factors to measure the true technical efficiency of the DMUs. They investigated the relationship between computed efficiency scores and a single performance measure of DMUs by using decision tree (DT) analysis. Using the relative efficiency scores of 57 DMUs, the results indicated that 2 out of the 11 models are the most appropriate to measure relative efficiency when intermediary approach was considered to evaluate the banks (Jain et al., 2016b). Da et al. (2018) assessed the efficiency of peripheral European domestic banks in order to examine the effects of bank-risk determinants on their performance using 2007– 2014. The authors applied both Data Envelopment Analysis and a Truncated Regression and suggested that financial ratios such as liquidity and credit risk were negatively affecting banks productivity, whereas capital and profit risk had a positive impact on their performance (Da et al., 2018). Kwon & Lee (2015) implemented a two-stage DEA and Neural Network and suggested that the effectiveness of an integrated performance model was empirically supported by its practical application to the financial banking operations across U.S. banks.

Wu et al. (2006) and Wu (2006 and 2009) in separate studies applied both DEA and Machine

Learning Algorithm with other models to make classifications and predictions. In the case of Wu (2006), the author combined DT with a two-stage DEA model in a study to detect the IT impact on 36 firms performance. The authors used Wang et al. (1997) dataset and categorized both the CCR and BCC DEA scores of the various DMUs into four classes under two separate scenarios. Using a hybrid model that consist of two modules, Wu (2006) and cited by Chen (2016) and Santos et al. (2017), DT model was able to predict the classes originally given to each DMU per the CCR and BCC scores at an accuracy of 69.44% and 64% respectively for the different “cutoff point” for efficient DMU. The author finally concluded that DT was easy to use and understood and suggested comparison studies of DEA-DT and DEA-neural networks (Wu et al., 2006). However, for the evaluation of the DEA scores, the author did not consider the efficiency score at each stage of the two – stage DEA model.

In another similar study, Delen et al. (2013) in determining the firm performance using a set of financial ratios adopted a two-step analysis methodology: first, using exploratory factor analysis (EFA) to recognized and confirmed fundamental dimensions of the financial ratios, followed a predictive modeling approaches to determine the possible associations between the firm performance and financial ratios. The authors tested their dataset using four popular decision tree algorithms (CHAID, C5.0, QUEST and CART) and performed information fusion-based sensitivity analyses to measure the relative importance of predictor variables. Delen et al. (2013) suggested that the CHAID and C5.0 decision tree algorithms produced the best prediction accuracy. Emrouznejad & Anouze (2010) proposed a framework to combine DEA with C&R for measuring the efficiency and productivity of DMUs. The authors suggested a set of rules that can be used by policy makers to discover reasons behind efficient and inefficient DMUs. Hamad & Anouze (2015) in a study presented a three-stage combined system comprising DEA, RF, and Logistic Regression to assess and forecast the effect of environmental factors on banks' efficiency. The model recognized five significant factors and their impacts on bank performance when illustrated with 151 banks in the Middle East and

North African (MENA) countries from 2008-2010. Hamad & Anouze (2015) utilized these variables to study their relationship with banks efficiency through logistic model. The authors concluded that the proposed model can be utilized by bank directors, and other stakeholders to better manage their banks.

Grzybowska & Karwański (2014) exhibited a technique for variable choice in light of Random Forest and gradient boosting approach and its application to companies ranking in DEA technique. The study recommended that application of RF methods appears to be a good method to variable choice for the needs of DEA. Finally they suggested that Random forests and gradient boosting can be relied upon to enhance the automation of procedures to evaluate the status of companies by banks and other financial institutions. Grzybowska & Karwański (2014) demonstrated that random forests yield higher outcomes in terms of both efficiency and expectation precision compared to other ensemble procedures. Using a recent sample of large corporate failures in the United States, Premachandra et al. (2009) examined the capability of DEA in assessing corporate bankruptcy by comparing it with logistic regression (LR). This study suggested that DEA outperforms LR in evaluating bankruptcy out-of-sample.

Sreekumar & Mahapatra (2011) presented an integrated approach by combining Data Envelopment Analysis (DEA) and Neural Network (NN) for assessment and prediction of Indian B-schools' performance. The study suggested that with a total of 49 Indian B-schools chosen for benchmarking purpose. The average score of efficiency was 0.625 with a standard deviation of 0.175 when Charnes, Cooper and Rhodes (CCR) model is used. Similarly, when the Banker, Charnes and Cooper (BCC) model was used the average score was 0.888 with a standard deviation of 0.063. According to the authors, the work offers a modest but inclusive methodology for improving performance of B-schools in India. Anthanassopoulos & Curram (1996) in a comparative analysis study to assess the efficiency of DMUs also used DEA and NN. The results of the study suggested that, irrespective of the differences between the two

models, both methods offer a useful range of information regarding the assessment of performance.

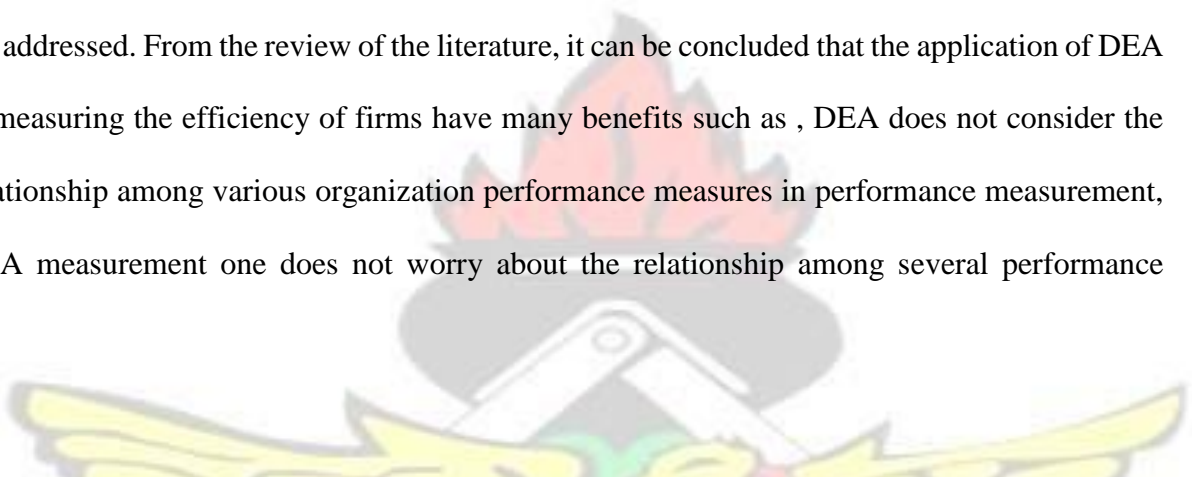
In attempt to assess the efficiencies of personnel, Azadeh et al. (2011) suggested that personnel specifications have greatest impact on total efficiency and proposed an algorithm for such a task. The proposed algorithm assesses the impact of personnel efficiency attributes on total efficiency through Data Envelopment Analysis (DEA), Artificial Neural Network (ANN) and Rough Set Theory (RST). According to Azadeh et al. (2011), the proposed algorithm was superior to the conventional and existing models and algorithms. Dash et al. (2006) combined data envelopment analysis (DEA) and Neural Networks (NNs) to assess the relative branch efficiency of Canadian bank. Dash et al. (2006) compared the results with the ordinary DEA results. The authors finally concluded that the two are analogous. Saher et al. (2019) in a study determined the firm and sector efficiency using data envelopment analysis for 121 listed firms, from 2004 to 2016. Saher et al. (2019) concluded that all firms were not equally efficient. The study also used a Logit/ Probit Regression model, and the results indicated that the brand value and type of sector has a positive impact on firm efficiency (Saher et al., 2019).

In conclusion, the study made a search for DEA-related studies in popular academic databases like Taylor and Francis, Elsevier, IEEE, Scopus, EBSCO and Science Direct. The results showed more than 1000 studies. These numbers of studies are impressive and confirm the fact that the issues regarding DEA application at measuring efficiency are still on the agenda and require the further investigation. DEA is widely used in efficiency measurement studies (Ascarya, 2007; Yumanita et al., 2008; Emmanuel, 1999; Grmanová & Ivanová, 2018; Halkos & Salamouris, 2004; Havidz & Setiawan, 2015; Jemric & Vujcic, 2002; Kamarudin, Sufian, Loong, & Anwar, 2017; Lampe & Hilgers, 2015; LaPlante & Paradi, 2014; Maletić, Kreća, & Maletić, 2013; Nand & Archana, 2015; Sarifuddin, Ismail, & Kumaran, 2015; Shibu & Ayekpam, 2018; Sreekumar & Mahapatra, 2011; Sufian et al., 2016 and Titko et al., 2014). However, studies on bank efficiency with DEA techniques in developing countries are limited particularly in Africa. This also means that the methodological

and informational gap in studies on bank efficiency conducted by local researchers using DEA should be an important issue. From the literature review also, most two-stage DEA model did not consider the efficiency score in each of the stages as an input variable in the next stage. Conceptually, the author consider outputs from organizations as a result of a sequences of important events where resources are applied at all the two stages as in the scenarios used by previous authors (Chen & Zhu, 2004; Wang et al., 1997). Thus, for the banks to have enough funds to invest in the second stage will depend on how efficient they are in collecting deposit. The current study uses a two-stage DEA concept that explicitly consider the efficiency score in all the stages i.e. both stages I and II as. This is acknowledged by using the efficiency of stage I and the total deposit that was realized at stage I as the two main inputs for stage II and its outputs as Percentage of Performing Loans Greenidge & Grosvenor (2010) and Profit. Finally, the overall efficiency was calculated by using the efficiency of stage II, Fixed Asset, IT budget and the Number of Employees as inputs and Percentage of Performing Loans and Profit accrued from investing the deposits in securities as the two main outputs of the overall stage. This proposed model does not only analyze the IT impact on the two stages, but also ensures that efficiency that was calculated at each stage is also used as inputs.

2.4 CHAPTER SUMMARY

This Chapter reviews DEA, Machine Learning Algorithms and their applications either individually or a combined DEA and Machine Learning Algorithm. For the DEA, the Chapter reviews DEA in general including background information, and applications of DEA in efficiency measurements. With regard to background information section, the definition of DEA, its data structures, algorithms used in DEA, advantages and disadvantages of DEA as well as major application of DEA for efficiency measurements particular in the business area are addressed. From the review of the literature, it can be concluded that the application of DEA in measuring the efficiency of firms have many benefits such as , DEA does not consider the relationship among various organization performance measures in performance measurement, DEA measurement one does not worry about the relationship among several performance



measures and DEA is an improved method to organize and analyze data since it allows efficiency to change over time (Lampe & Hilgers, 2014). It also does not involve any previous postulation on the requirement of the best practice frontier. However, the application of only pure DEA and its models suffer from weak discrimination power Ashoor (2012) and sensitive to the presence of outliers and statistical noise (Dash et al., 2006). It is also very difficult to use only DEA to predict the efficiency and performance of other or new Decision Making Units (LaPlante & Paradi, 2014; Peter Wanke et al., 2015).

The second part of the Chapter Two also focuses on Machine Learning Algorithms its applications and previous studies on combined DEA and Machine Algorithms especially in the financial industry. In background information aspect, the origin and definition of Machine Learning Algorithms and types of Machine Learning Algorithms are discussed. It also introduces the topmost/most commonly used Machine Learning Algorithms and evaluation or performance measurements of Machine Learning Algorithms. Based on the review, four among the ten topmost Machine Learning Algorithms namely, Decision Tree, Random Forest, Neural Network and Logistic Regression were extensively discussed in this chapter and were used for the study. Finally application of Machine Learning Algorithms whether as a single tool or combined with others showing different cases studies are presented. From the review of the literature on Machine Learning Algorithms, it can be suggested that a lot of Machine Learning

Algorithms have been developed and utilized in the business and financial sectors for predictions and forecasting. Nevertheless, little attention have been given to prediction and classification of bank branches using their efficiency scores across developing countries Mohd (2001) especially using a combined two-stage DEA and Machine Learning Algorithms, where efficiency score at each stage is also used as an input for the next stage. With this, we can have a study that combines a two -stage DEA model with different Machine Learning Algorithms, where the efficiency at each stage is also considered as input for the next stage using dataset from a developing country.



CHAPTER 3

METHODOLOGY

3.0 Introduction

This chapter discusses the research methodology and the methods used in carrying out the study. The four Machine Learning Algorithms namely; Decision Tree, Random Forest Artificial Neural Networks and Logistic Regression and the four proposed models that were

realized by combining DEA with Machine Learning Algorithm is used for making predictions in a greater part of this thesis. These four proposed models are: Data Envelopment Analysis-Decision Tree (DEA-DT), Data Envelopment Analysis-Random Forest (DEA-RF), Data Envelopment Analysis-Neural Network (DEA-NN) and Data Envelopment Analysis-Logistic Regression (DEA-LR).

3.1 DATA

3.1.1 Study Area and Data Description

The main case study area of the study is the Ghanaian banking sector. The study was undertaken basically to study efficiency of Ghanaian banks. Due to the large number of commercial bank branches operating in the country as of the time of this study 2016 (Ghana, 2017), the study was limited to 33% of the total commercial (universal banks) branches in Ghana. These banks were mostly in Greater Accra, Ashanti, Western, Brong Ahafo, Eastern, Northern, Upper East, Volta and Central Regions. The study was limited to these areas because of the availability of information and logistics. Data for the study was also the audited 2016 financial statements from the various bank branches. The map of the study area is shown below in Figure 3.1.

The Study Map Area

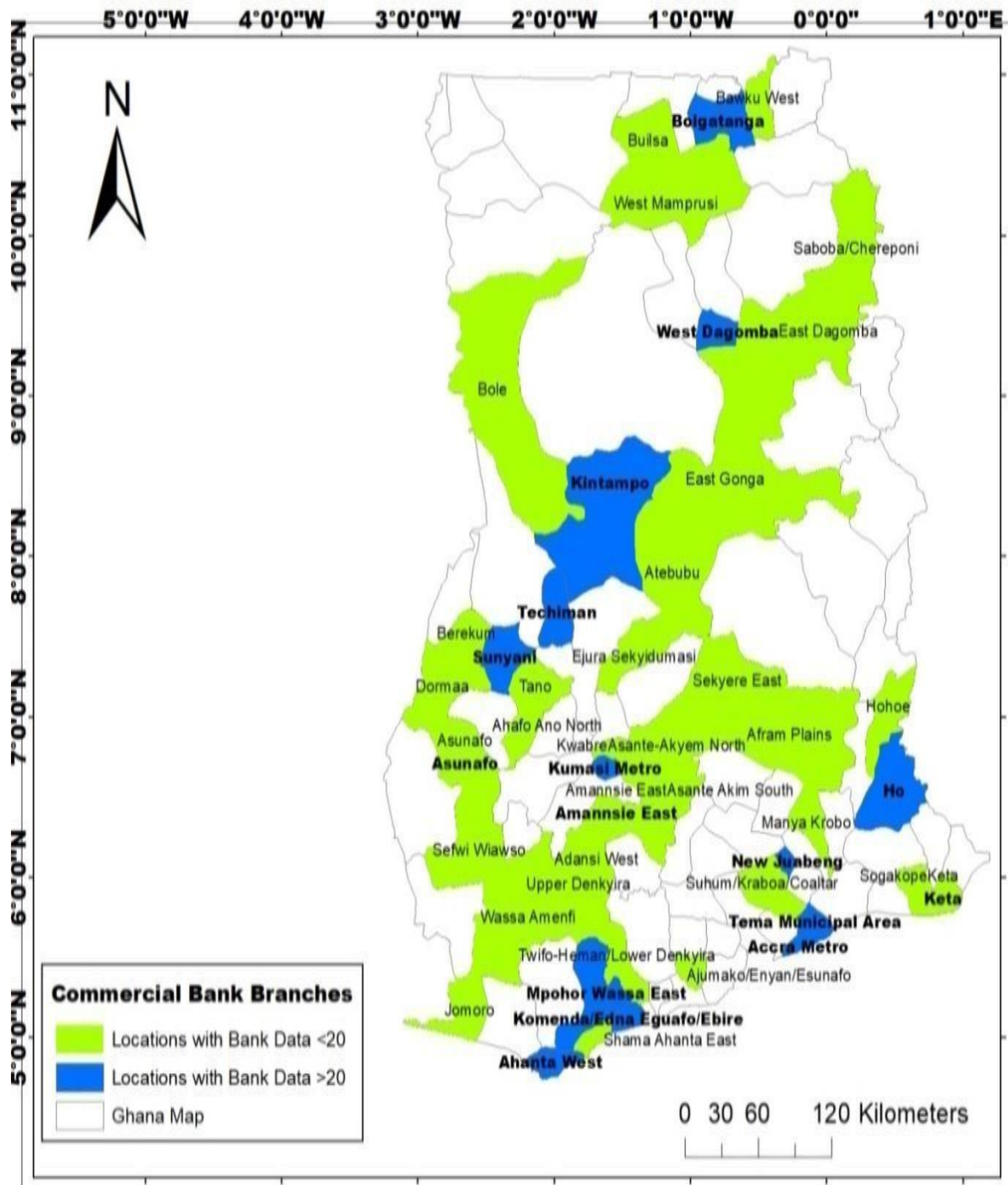


Figure 3.1: Map of the case study area showing the distribution of bank branches used for the study (Author's construct).

3.1.2 Data Collection and Sample Size

Arrangement was made and the necessary protocols were followed to collect data (Fixed Assets excluding IT, Number of Employees, IT expenditure, total deposit, percentage of performing loans and profits accrued from investing deposit) from the selected banks branches using their 2016 audited financial statement. The sample size for the study data was calculated using minimum sample size formula in equation (10) at confidence level of 95% and 5% margin of error. This formula was used because the study is a cross-sectional survey and the response variable is also qualitative (Cochran, 1977) .

$$\text{Sample size (n)} = \frac{2 (Z_{\alpha/2})^2 \cdot P(1-P)}{E^2} = \frac{(1.96)^2 \cdot (0.5)(0.5)}{(0.05)^2} = 367.32 \approx 367$$

(44)

Where α -level of significance which in this case was selected to be 5%, meaning ($Z_{\alpha/2}$) = 1.96 at (95% confidence level). P is the estimated population proportion which was chosen to be 0.5. This is the possessing attribute of interest based on previous studies or pilot studies. If no approximation of P is known, P=0.5 is used which will give a sample size sufficiently large to guarantee accurate results (Bluman, 2009). Based on this minimum sample size value of 367, requests were made to collect data from about 650 universal bank branches in the selected study area and follow-ups were made to explain the project to the banks and also to have some firsthand information about the operation of the banks.

A total of 472 requests were granted (thus 472 bank branches data were supplied given a response rate of 73%). After analyzing and cleaning the data, the total observation of the dataset came to 444 which were used for the study. This number is also consistent with Charter (1999) and cited recently by (Akena et al., 2018), that the total number for generalizability and reliability research studies is fairly individualistic but a minimum of 400 observations is recommended.

The Cedi value of IT budget data, data on the percentage of performing loans on the various bank branches, total number of employees at each branch, profit accrued from the banks investments and the Cedi value of the total deposit were obtained from the various branches. This study is different from previous works by Wang et al. (1997) and Wu (2006) which was cited by Chen (2016) and Santos et al. (2017) in the sense that the study used the percentage of performing loans (non-performing loans rate minus 100%). The data in Appendix A shows the complete set of 444 observations which gave us the dataset for this study. This data sample

contains 444 observations on DMUs which is about 33% (comprising 17 universal banks) of the total commercial bank branches in Ghana.

3.2 THE PROPOSED BANK EFFICIENCY PREDICTION FRAMEWORK

The framework suggested in this study which was used to build the predictive models consists of three different stages, data collection stage, data preprocessing stage and the predictor development model stage (Decision Tree/Random Forest Classifier model). This means that the dataset for the model development goes through three different stages. It was used to build the predictive models used for predicting the efficiency of banks. Figure 3.2 is the proposed the framework used for the entire study. The framework also has a Neural Network version which was also used for developing the NN predictive models for the study.

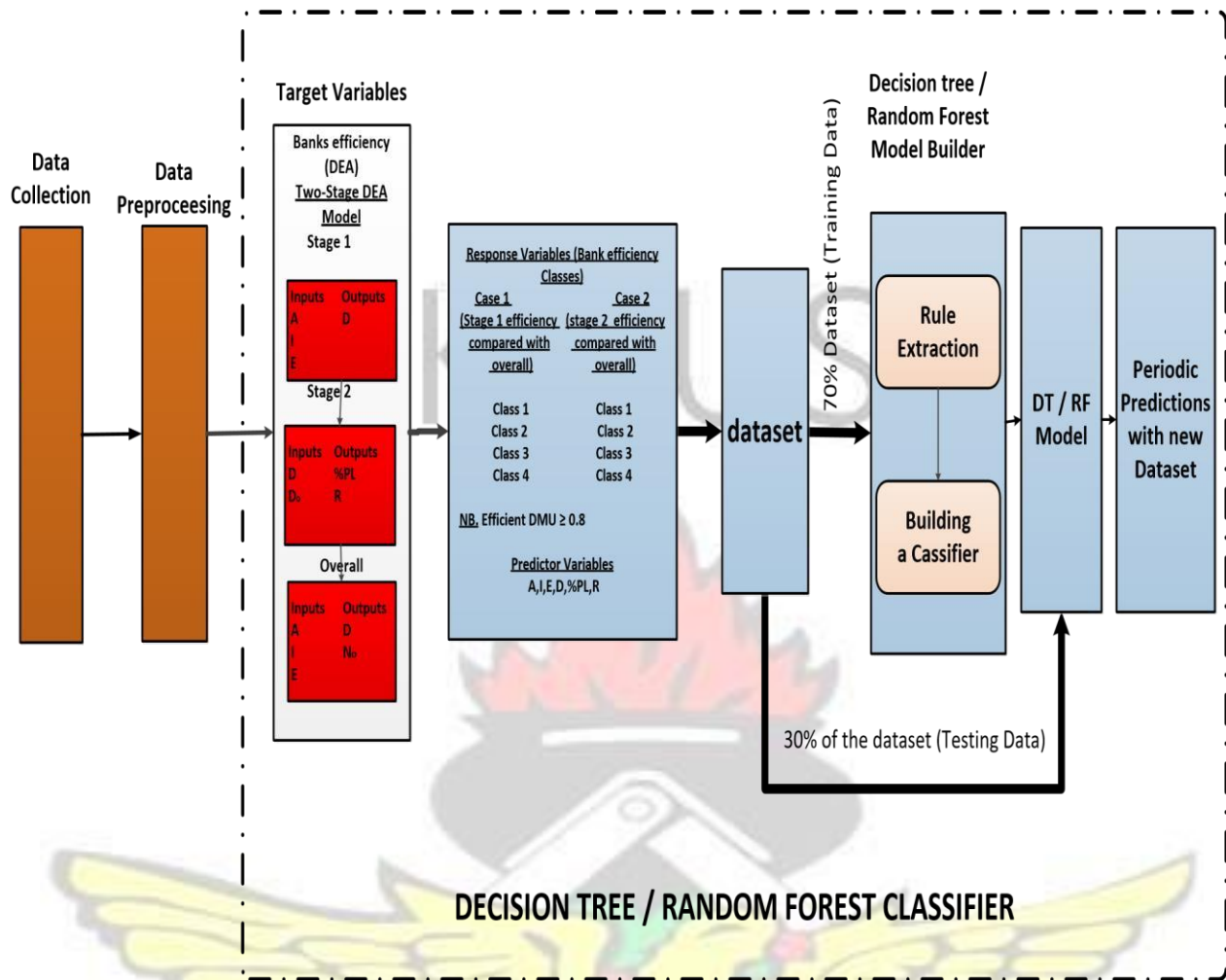


Figure 3.2: The Banks efficiency prediction framework (Author's construct)

3.2.1 Dataset for the Model Development

3.2.1.1 Determining the Response Variables (Bank Efficiency Scores and Classes)

The response variables (efficiency scores of the bank branches) of the study were determined using a two-stage Data Envelopment Analysis (DEA).

The Basics of DEA

Data Envelopment Analysis is a non-parametric method that produces a comparative ratio of weighted outputs to inputs for each Decision Making Unit (DMU) under consideration, i.e. relative efficiency score (Banker et al., 1984 and Cooper et al., 1978). This score is normally recorded as a figure between the percentages of 0-100% or a unit that falls between 0-1. A

DMU with a score less than 100% or less 1 unit is classified as ineffective comparable to similar units in the sample (Avkiran, 2006). According to Wang et al. (1997), it is a scientific programming method used to compute the relative specialized efficiencies of gainful units.

The DMUs employ several inputs to generate several outputs where the inputs cost and output cost are assumed not to be known. The framework was formerly used to examine the relative efficiency of non-profit organizations which was quickly followed by other profit making organization such as financial institutions, hospitals, the US Air Force, airports, schools and courts because of its strength (Avkiran, 2006).

For instance, Coelli & Rao (2005); Jagoda et al. (2013); Lusigi (1997); Oeij, Looze, Have, Rhijn (2011) and Syverson (2011) used this model in their various studies. The model shown in Figure 3.3 was originally suggested by Charnes et al. (1978) , CCR which have seen several extensions including the one by Banker et al. (1984) , BCC which is considered a significant extension (Wang et al., 1997). The CCR, Constant Returns to Scale, CRS Charnes et al. (1978) was adopted for the study in reference to other similar models.

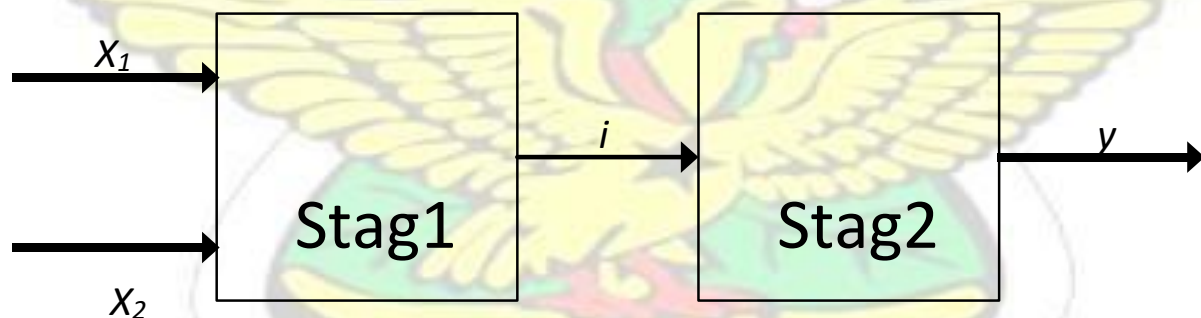


Figure 3.3: The Classical two-stage DEA model taking two (2) inputs adapted from Wu (2006).

Based on data matrix $(X; Y)$, one can calculate the input oriented efficiency of each observation “o” (in this case DMU) by calculating n number of times the subsequent linear programming problem proposed by Charnes et al. (1978) and popularly known as *DEA CCR* model:

$$\min \theta$$

$$\theta, \lambda$$

Subject to

$$\theta x_o \geq X \lambda$$

$$Y \lambda \geq y_o$$

$$\lambda \geq 0.$$

The optimum answer to this program is based on CRS technology denoted by θ_{CRS}^* .

The restrictions entail the DMU $(\theta_{CRS} x_o, y_o)$ to belong to P_{CRS} , while the purpose is to pursue the minimum θ_{CRS} that decreases the input vector x_o radially to $\theta_{CRS} x_o$ while still in P_{CRS} . A possible answer signaling radially efficiency is $\theta_{CRS}^* = 1$, therefore if $\theta_{CRS}^* < 1$, the DMU is radially inefficient and $(\lambda X, \lambda Y)$ surpasses (x_o, y_o) .

With regard to this property, we express the additional input excesses and output shortfalls by the following slack vectors: $s^- \in \mathbb{R}^m$ and $s^+ \in \mathbb{R}^s$, respectively. Therefore: $s^- =$

$\theta_{CRS}^* x_o - X \lambda$, and $s^+ = Y \lambda - y_o$ with $s^- \geq 0$ and $s^+ \geq 0$ for any possible solution (θ, λ) . To get the probable input excesses and output correct, the second stage program that integrates the optimal value θ_{CRS}^* and fix radial inefficiency is calculated as:

$$\max_{\lambda, s^-, s^+} \quad \omega = e s^- + e s^+$$

Subject to:

$$s^- = \theta_{CRS}^* x_o - X \lambda$$

$$s^+ = Y \lambda - y_o$$

$$\lambda \geq 0, s^- \geq 0, s^+ \geq 0,$$

Where $e(1, \dots, 1)^T$ so $e s^- = \sum_{i=1}^m s_i^-$ and $e s^+ = \sum_{i=1}^s s_i^+$.

This means that a DMU is efficient technically on condition that the optimum solution

$(\theta_{CRS}^*, \lambda^*, s^{-*}, s^{+*})$ of the two above program satisfy $\theta_{CRS}^* = 1, s^{-*} = 0$ and $s^{+*} = 0$, so no

equi-proportional contraction of inputs, and individual inputs reduction and outputs growths are possible (Álvarez et al., 2016).

The analysis of technical efficiency assuming Variable Returns to Scale (VRS) proposed by Banker et al. (1984) and cited by Álvarez et al. (2016) called DEA BCC model, also consider this production possibility set $P_{VRS} = \{(x, y) | x \geq X \lambda, y \leq Y \lambda, e \lambda = 1, \lambda \geq 0\}$. Thus the only distinction with that of the CCR model (CRS) is the adjunction of the condition $\sum \lambda = 1$. Analyzing the Value Return to Scale (VRS) efficiency of each DMU under discussion alongside the succeeding second stage program similar to equation (12)

gives the equivalent optimum solution $(\theta^*, \lambda^*, s^{-*}, s^{+*})$. As stated earlier, a DMU is efficient according to the VRS technology if and only if the optimum solution of the two programs satisfy $\theta_{VRS} = 0$ and $s = 0$.

The DEA Model and Notation

As stated earlier, the operations of the various bank branches were considered to be a twostage operation. In this case, banks in Ghana use IT infrastructure, Number of staff and their fixed asset to collect money from their customers called deposit in the stage I. The Cedi value of the total deposit was used as the output variable while the Cedi equivalent of their IT (IT expenditure), Number of employees and fixed assets were used as the input variable. To evaluate the bank performance and efficiency, it is incumbent to align its resources to its fundamental business goals. For instance, IT facilities link Computers installed with needed software, mobile phones, Automated Teller Machines (ATMs), internet facilities etc. are vital tools that banks use to collect or gather deposit and also use it to invest their deposits into securities through electronic transfers and process loans for their customers. The study therefore, noted both stages I and II as resource-related value-added events and considered the efficiency of both stages as very important.

A bank's profit after all the various deductions including tax is centered on the revenues accrued from the investments and the loans given out in stage II. In this case it will be totally improper

to only assess a bank's performance and efficiency only on their profit margins without looking at how their loans given out are performing as regards percentage of performing loans. This is because a bank with weak percentage of performing loans stands a high probability of losing net profit and can also threaten the bank's existence. The study therefore denoted this situation by the variable, Percentage of Performing Loans.

The classical two-stage DEA for efficiency computing and analysis by Wang et al. (1997) and Wu (2006) cited by Chen (2016) and Santos et al. (2017) to compute the efficiency score of each DMU was adapted in this study. In this DEA model, the various units (banks) performance or efficiency measures were grouped into inputs and outputs. Using a bank as an example to derive the model, the business processes of the bank and activities is viewed as a two-stage, wherein the first stage (Deposits Stage) of the model consists of collection of funds (Deposits) in Ghana Cedis from customers using their fixed asset, number of workers (Employees at each unit) and IT facilities (thus IT expenditure). In the next stage (Investment Stage), the banks use the deposits accumulated in stage I to invest in securities and also give loans to their customers and returns (Profits) generated from the investment in securities and percentage of performing loans, which are a good indicators of risk status are used as two outputs in stage II. The various variables used in the proposed model are defined as follows:

Deposit Stage I Input:

- Fixed Assets (billions of GH) denoted as A
- Total IT expenditure (billions of GH) denoted as I
- Total number of Employees denoted as E

Investment Stage II Input:

- The efficiency of stage I denoted as D_o .
- The deposit also denoted as D.

Output:

Percentage of Performing loans (PL)

Profit accrued from investing in securities (R)

Input

Efficiency of stage II denoted as No

Fixed Assets (billions of GH) denoted as A

Total IT expenditure (billions of GH) denoted as I

Total number of Employees denoted as E

Output:

Percentage of Performing Loans (PL) .

Profit accrued from investing in securities (R)

After using the classical CCR model of DEA, the efficiency score for each DUM in each stage resulted in the three types of efficiency score as follows:

Efficiency of stage I, Do

Efficiency of stage II, No

Overall efficiency, G

DEA models used at the deposit stage, I produce stage I efficiency for each DUM and the DEA applied at the investment stage, II also gives stage II efficiency for each DMU. DEA models employed at overall stage finally give us the overall efficiency for each DMU. The proposed

Overall Stage

-
-
-

-

-

-

-

-

-

DEA model for the banks two-stage operation is depicted in the Figure 3.4. **The DEA Model for the Bank's Dual Role Operations**



Figure 3.4: The Proposed Dual Role DEA Model adopted from Appiahene et al. (2019)

Efficiency Calculations

For each DMU, the technical efficiencies in both stages and their corresponding overall efficiencies were analyzed using CCR DEA algorithm proposed by Mehrabiana (2013) which depended on non-Archimedean Charnes-Cooper-Rhodes (CCR) framework below.

Algorithm 3.1 : Two-Phase Algorithm BuildHull

Step 0:

Input problem data, $k = 1$.
 $flag(j) = 1$ if DMU_j classified, else 0.
 $flag(j) = 0, j = 1, \dots, n$.
 Compute an assurance value $\bar{\epsilon}$, $\bar{\epsilon} \leftarrow \bar{\epsilon}$.

Step 1:

Solve problem $E(P_k)$.
 case $(\theta_k < 1)$ if $((\lambda_k, \omega_k)$ is an SCSC pair),
 $DMU_k \in U_M$,
 if $(1^T S_k = 0)$, **if** $(\lambda^{k_k} > 0, \lambda^{\ell_j} = 0 \text{ for } j \neq k)$,
 $DMU_k \in NE$.
 else,
 $DMU_k \in NE'$,
 if $((\theta_k, \lambda_k, S_k)$ is unique & $(\lambda^{k_k} > 0, \lambda^{\ell_j} = 0 \text{ for some } \ell \neq k)$,
 $DMU_{\ell} \in E, flag(\ell) = 1$ **endif**
 else,

$DMU_k \in NF$,
 if $((\theta_k, \lambda_k, S_k)$ is unique & $(\lambda^{k_k} > 0, \lambda^{\ell_j} = 0$ for some $\ell \neq k)$,
 $DMU_\ell \in E$, $flag(\ell) = 1$.
 endif case $(\theta_k = 1)$ if
 $(1^T S_k = 0)$, if $(\lambda^{k_k} > 0)$,
 $DMU_k \in E$. else
 $DMU_k \in E'$. endif
 else, $DMU_k \in F$.
 endif

Step 2:

$k \leftarrow k + 1$ if $(k > n)$,
 stop.
 else, if $(flag(k)=1)$, go to Step 2
 else, go to Step 1.

This algorithm has its package Robust Data Envelopment Analysis (rDEA) Version 1.2 -5
 Simm (2016) and was implemented in R using R studio with screen shot shown in Figure 3.5
 and the results are shown in appendix B.

The efficiency of stage I was calculated using IT expenditure (GH ₵), Fixed Asset (GH₵) and
 the Number of employees at each DMU as inputs and deposit as the main output. Regarding
 stage II, the deposit (GH₵) realized from stage I and the efficiency value of stage I were used
 as inputs with bank profit and the percentage of performing loan as outputs.

The efficiency of the overall stage was also calculated using the efficiency value from stage II,
 Fixed assets, Number of employees and IT expenditure while their output was the banks profit
 and the percentage of performing loans.

The R Studio

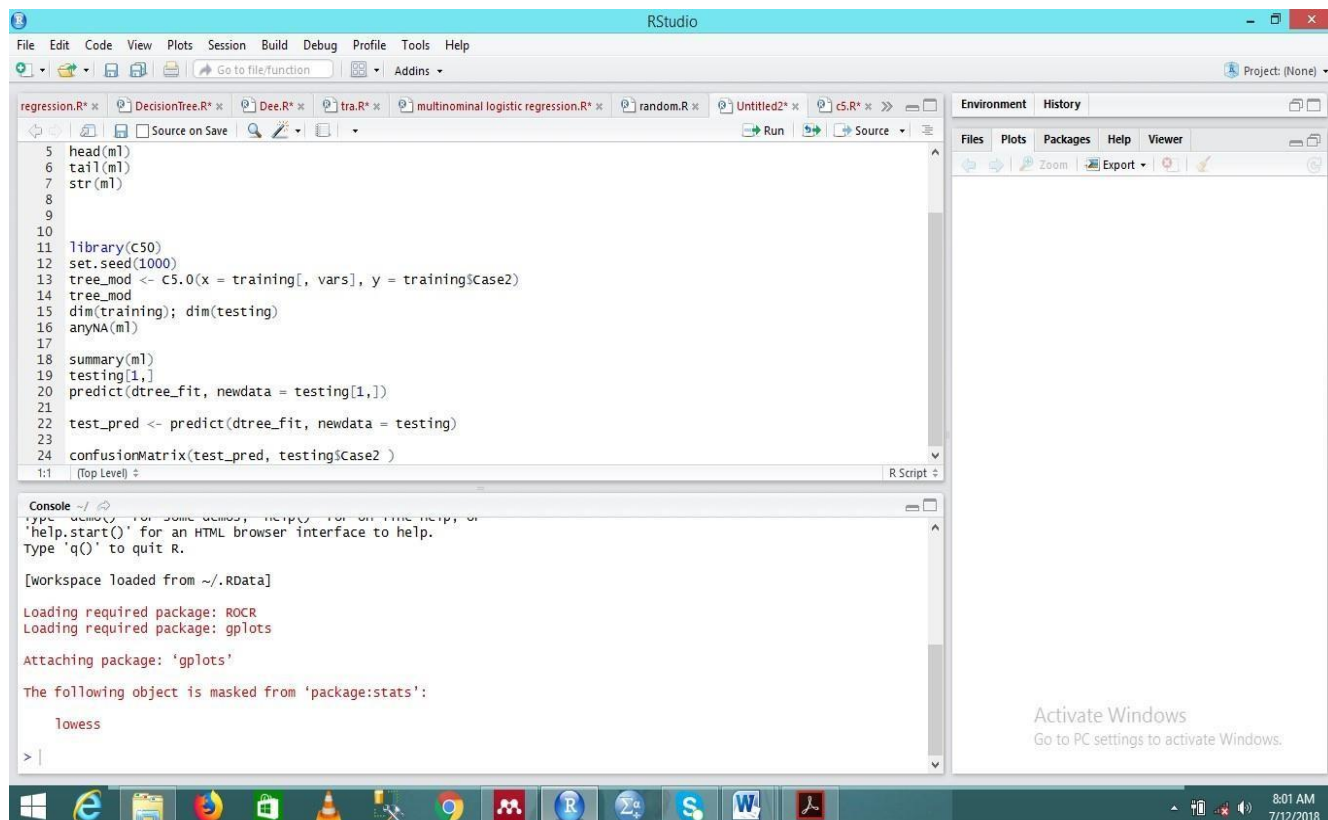


Figure 3.5: Screen shot of R studio used for the programming (Author's construct).

3.2.1.1.2 Classification of the Banks' Efficiencies Scores

Using the overall efficiency as the basis, the efficiency of both stage I (banks efficiency in collecting deposit from customers) and the stage II (the banks efficiency in investing their deposit into securities and given out loans) were categorized into classes as was done by Wang et al. (1997) and Wu (2006) cited by Chen (2016) and Santos et al. (2017) using the proposed Bank Classification Algorithm, BC Algorithm (Appiahene & Missah, 2020) below:

The Proposed Bank Classification Algorithm (BC Algorithm)

- *Let DMU represent Bank's*
- *Let A represent Bank's fixed Asset excluding their IT*
- *Let I represent Bank's IT expenditure*
- *Let E represent Bank's total number of Employees*

- Let D represent Bank's total deposits
- Let Do represent Bank's efficiency score in collecting deposit from customers
- Let No represent Bank's efficiency score in investing the deposits collected from customers

Let G represent Bank's overall efficiency score in doing business

Let PL represent Bank's percentage of performing loans

Let R represent Bank's profits accrued from investing their deposit.

Let Efficiency Bank be a Bank with efficiency score ≥ 0.8 comparative to similar units in the sample

// calculating the Efficiency of the Banks in the Deposit stage (Do), Investment stage (No) and their corresponding Overall stage (G)

1. Start
2. Input A, I, E, D, PL, R of a DMU
3. Use A, I and E as input variables and D as output variables and determine the DMU's Do using the **CCR Two-Phase DEA BuildHull Algorithm**
4. Use Do and D as input variables and PL and R as output variables and determine the DMU's No using the algorithm suggested in **STEP 8**.
5. Use A, I, E and No as input variables and PL and R as output variables to determine DMU's G using the **STEP 8** algorithm.

// Classifying the Banks based on their Do, No scores

// Categorizing the Banks based on their Do on the basis on their overall efficiency

- Let **Class 1** be Banks efficient in collecting deposit but was not able to attain overall efficiency.

- Let **Class 2** be Banks that realized overall efficiency even though they were not efficient in collecting deposit.
- Let **Class 3** be Banks that were efficient collecting deposit and still had overall efficiency.

6. If DMU's $Do \geq 0.8$ & $G < 0.8$

Class 1 <----- DMU

else

If DMU's $G \geq 0.8$ & $Do < 0.8$

Class 2 <----- DMU

Else

If DMU's $Do \geq 0.8$ & $G \geq 0.8$

Class 3 <----- DMU

else

Class 4 <----- DMU

// categorizing the Banks based on their No on the basis on their overall efficiency

Let **Class 1** be Banks efficient in investing deposit but was not able to attain overall efficiency.

Let **Class 2** be Banks that realized overall efficiency even though they were not efficient in investing deposit.

Let **Class 3** be Banks that were efficient in investing deposit and still had overall efficiency.

- Let **Class 4** be Banks inefficient in collecting deposit and cannot also achieve overall.

- -
 -
 - *Let **Class 4** be Banks inefficient in investing deposit and cannot also achieve overall efficiency*
7. *If DMU's $No \geq 0.8 \&\& G < 0.8$*
***Class 1** <----- DMU else*
- If DMU's $G \geq 0.8 \&\& No < 0.8$*
***Class 2** <----- DMU*
else
- If DMU's $No \geq 0.8 \&\& G \geq 0.8$*
***Class 3** <----- DMU*
else
- Class 4** <----- DMU*
8. *End*

3.2.1.2 Predictor Variables

These are variables also called independent variables or experimental variables employed in statistical analysis to forecast or predict another variable called target or dependent variable (Chi et al., 2011; El-habil, 2012 and Tsai et al., 2009). For this study, the predictor variables were: Fixed Assets excluding IT, IT expenditure and Number of Employees.

3.3 THE COMPUTER PROGRAMMING PLATFORM

The various computer programming presented in this thesis were done using the RMiner. The R can be regarded as a computer programming language with a huge inbuilt library of predefined functions that can be utilized to implement several jobs (TutorialPoint, 2016). R allows for the incorporation of codes written in the C, C++, .Net, Python or FORTRAN languages for efficiency. R is open source software accessible under the GNU General Public License, and has pre-compiled binary versions which are provided for various operating systems like Linux, Windows and Mac. The RMiner offers a set of comprehensible functions (e.g. mining, saveMining) for classification and regression tasks. In particular, the library has the caret, C5.0, randomForest, neuralnet and kernlab (SVM) packages. It is important to mention that all the programming codes used for the entire study was code in R using the R studio. The codes were run on an intel® Celeron® CPU N2840 @ 2.16GHz with an installed memory (RAM) of 2.00GB. The machine is also a 64-bit operating system, x64 based processor with windows 8.

3.4 BUILDING THE PREDICTIVE MODELS FOR PREDICTING BANKS EFFICIENCIES

3.4.1.1 The C5.0 Algorithm

The C5.0 Decision Tree algorithm is an extension or improvement of both ID3 and C4.5 algorithms, and therefore shares the same basic principles with the previous algorithms ID3 and C4.5 (Ananda & Wibisono, 2014). This algorithm for tree construction uses the greedy algorithm Ross (1994) and Yuan & Lin (2006) principle that construct tree from top to down (top down) recursively by divide and conquer (Ananda & Wibisono, 2014). The C5.0 algorithm is shown in algorithm 3.2.

Algorithm 3.2 : *Generate_decision_tree*

Input: data training samples; list of attributes; attributes_selection_method.

Output: decision tree.

Method:

1. *create a node N,*
2. *if samples has the same class, C, then*
3. *return N as leaf node with class C label;*
4. *If list of attributes is empty then*
5. *return N as leaf node with class label that is the most class in the samples.*
6. *Choose test-attribute, that has the most GainRatio using attributes_selection_method;*
 give node N with test_attribute label;
 for each a_i pada test-attribute;
 Add branch in node N to test-attribute = a_i ;
- 7.
- 8.
- 9.
10. *Make partition for sample s_i from samples where test-attribute = a_i ;*
11. *If s_i is empty then*
12. *attach leaf node with the most class in samples;*
13. *else attach node that generate by Generate_decision_tre*

$(s_i, \text{attribute-list}, \text{test-attribute});$

14. **endfor**

15. **return** $N;$

Using the C4.5 algorithm as shown above, a DT model can be constructed as follows. For instance, if there is single set of training dataset S that consist of the attributes $(P1, P2, P3, \dots)$ and classes also comprising of $(M1, M2, M3, \dots)$. The C4.5 algorithm would work as follows:

1. If S is not free and all the dataset samples contain the equal class of M_i , then the Decision Tree for S is a leaf node with label M_i .
2. Else if the attribute is free then the Decision Tree has a leaf node with label M_j where M_j is the uppermost class in the training samples S .
3. If S comprises a sample that has a dissimilar class of the partition S into $S_1, S_2, S_3, \dots S_n$. Training sample S partitioned by different values of attribute P_m , which at the time assumes the role of a parent node. Suppose P_m involves 3 categories of values that are n_1, n_2, n_3 , then S will be partitioned into three subsets, namely the value of $P_m = n_1, n_2 = P_m$, and $P_m = n_3$. This process continues recursively with the base case of step 1 and step 2. Attribute that would serve as the parent node or attribute that would partition the data is done by calculating the gain. Gain is used to select the attributes to be tested based on information theory concepts of entropy.

Entropy: Information entropy (or Shannon Entropy) defined by Shannon (1948) is an assessment which is founded on the possibility that is employed to calculate the quantity of doubt (Ananda & Wibisono, 2014; Olaru & Wehenkel, 2003 and Pandya & Pandya, 2015).

Let S be a random variable with number

$s_i, \quad i \in \{1, \dots, n\}, \quad \text{and probability mass function } p.$

$$\text{Entropy of } S, H(S) = \sum_{i=1}^n -p(s_i) \log_2(p(s_i)) \quad (45)$$

Note: For the purpose of entropy calculation, $\log_2(0) = 0$

The entropy function $H(S)$ in the above formulae satisfies the following:

- $H(S)$ is continuous in $p(s_i), i \in \{1, \dots, n\}$.
- If $p(s_i) = \frac{1}{n}$ for $i = 1, \dots, n$, then $H(S)$ is monotonically increasing function.
- $H(S, Q) = H(S) + H(Q)$ is additive for any choice of two independent random variables S and Q .

Information Gain

Information gain in C5.0 algorithm is the variation in entropy that ensues just when after segregating the data based on an attribute. With an understanding of a random variable S , one can suggest that there are N attributes $\{a_i\}_{i=1}^N$ connected to each observation, where $a_i = a_i(S)$ for all $i = 1, \dots, N$. This suggests observations on the form $(s_i, a_i(s_i), \dots, a_N(s_i))$, where each attribute a_i is in itself a realization of a random variable.

Assuming $\{S_i\}_{i=1}^n$ is the set of possible outcomes of the random variable S with entropy $H(S)$.

Let A be the set of all possible values of the N -tuple $(a_i(s_i), \dots, a_N(s_i)), i = 1, \dots$

, n . For each function a_i , let \mathcal{A}_i denote the set of possible outcomes of a_i .

For a set of realizations S of the random variable S and attributes A , then the information gain

a split of S on attribute $a_i(s_i)$ is given by $IG(S, \mathcal{A}_i) = H(S) - \sum_{\alpha \in \mathcal{A}_i} \frac{|S_\alpha|}{|S|} H(S_\alpha)$ (46)

Where $S_\alpha = \{s \in S | a_i(s_i) = \alpha\}$ (47)

3.4.2 The Random Forest (RF) Algorithm

Algorithm 3.3: Binary Recursive Partitioning

Let $\mathcal{D} = \{(x_1, y_1), \dots, (x_N, y_N)\}$ represent the training data, with $x_i =$

$(x_{i,1}, \dots, x_{i,p})$.

1. Begin with all observations $\{(x_1, y_1), \dots, (x_N, y_N)\}$ in a single node.
2. Reiterate the following steps recursively for each unsplit node until the ending condition is met:

- a. Discover the finest binary split amongst all binary splits on all p predictors.
 - b. Split the node into two offspring nodes using the best split (step 2a).
3. At each prediction at x , forward x down the tree till it settles in a terminal node. Let k denote the terminal node and let y_{k1}, \dots, y_{kn} represent the target variables of the training data in node k . Predicted values of the target variable are
 - $\hat{h}(x) = \bar{y}_k = \frac{1}{n} \sum_{i=1}^n y_{ki}$ for regression
 - $\hat{h}(x) = \underset{y}{\operatorname{argmax}} \sum_{i=1}^n I(y_{ki} = y)$ for classification where $I(y_{ki} = y) = 1$ if $y_{ki} = y$ and 0 otherwise

Algorithm 3.4 Random Forest

- Let $\mathcal{D} = \{(x_1, y_1), \dots, (x_N, y_N)\}$ represent the training dataset, with $x_i = (x_{i1}, \dots, x_{ip})$. For $j = 1$ to J
1. Select a bootstrap sample \mathcal{D}_j of size N from \mathcal{D} .
 2. Utilizing the bootstrap sample \mathcal{D}_j in step(2) as the training dataset, fit a tree by using binary recursive partitioning :
 - a. Begin with all observations $\{(x_1, y_1), \dots, (x_N, y_N)$ in a single node.
 - b. Reiterate the following steps recursively for each unsplit node till the ending condition is met:
 - i. Pick m predictors at random from the p available predictors.
 - ii. Discover the finest binary split amongst all binary splits on all m predictors
 - iii. Split the node into two offspring nodes using the split from step ii. To make a prediction at a new point x

- $\mathcal{F}(\mathbf{x}) = \frac{1}{J} \sum_{j=1}^J \hat{h}_j(\mathbf{x})$ for regression
- $\mathcal{F}(\mathbf{x}) = \underset{y}{\operatorname{argmax}} \sum_{j=1}^J I(\hat{h}_j(\mathbf{x}) = y)$ for classification where $\hat{h}_j(\mathbf{x})$ is the prediction of the target variables at \mathbf{x} using the j th tree.

Utilizing the Out-Of-Bag Data

It is most likely that when taking a bootstrap sample from the dataset as illustrated in Algorithm 2.8 above, some observations would not find their way into the bootstrap sample. They are denoted as are “out-of-bag data”, and may be very important in approximating generality error and predictor variable significant or importance. To approximate generality error, it is important to mention that if the trees are enormous, predictions trustingly attained by means of all the trees would be excessively positive if they are used to predict the target variable for observations that were in the training set D . Because of this, prediction of the target variable for observations found in the training dataset is accomplished by utilizing trees whose observation is out-of-bag. These types of predictions are referred to as out-of-bag predictions as illustrated in algorithm 3.5.

Algorithm 3.5 Out-of-bag predictions

Let

\mathcal{D}_j denote the j th bootstrap dataset and $\hat{h}_j(\mathbf{x})$ represent the prediction at

\mathbf{x} from the j th tree, for $j=1, \dots, J$. For $i=1$ to N :

1. Let $\mathcal{g}_i = \{j: (\mathbf{x}_i, y_i) \notin \mathcal{D}_j\}$ and let j_i be the cardinality of \mathcal{g}_i
2. State the out-out-of bag prediction at \mathbf{x}_i to be

- $\mathcal{F}_{oob}(\mathbf{x}_i) = \frac{1}{j_i} \sum_{j \in \mathcal{g}_i} \hat{h}_j(\mathbf{x}_i)$ for regression
 - $\mathcal{F}_{oob}(\mathbf{x}_i) = \underset{y}{\operatorname{argmax}} \sum_{j \in \mathcal{g}_i} I(\hat{h}_j(\mathbf{x}_i) = y)$ for classification where $\hat{h}_j(\mathbf{x})$ is the prediction of the target variables at \mathbf{x} using the j th tree
-

In the case of regression with squared error loss, generality error is normally approximated through the out-of-bag mean squared error (MSE):

$$MSE_{oob} = \frac{1}{N} \sum_{i=1}^N (y_i - \mathcal{F}_{oob}(x_i))^2 \quad (48)$$

where $\mathcal{F}_{oob}(x_i)$ is the out-of-bag prediction for observation i . In the case of classification with zero one loss, generality error is also estimated by applying out-of-bag error rate:

$$E_{oob} = \frac{1}{N} \sum_{i=1}^N I_N(y_i \neq \mathcal{F}_{oob}(x_i)) \quad (49)$$

Most misconceptions could be that the out-of-bag error rate is attained by calculating the outofbag error rate for individual tree, and finding the mean error rates to suggest the out-ofbag error rate for the forest Cutler et al. (2011). Rather, we apply the error rate of the out-ofbag predictions. This gives us an opportunity to find a class-wise error rate for individual class, and an out-of-bag “confusion matrix” Cutler et al. (2011); Cutler (2010); Gislason et al. (2006) by cross-tabulating y_i and $\mathcal{F}_{oob}(x_i)$.

Permutation Variable Importance

Assessment of the most important and significant predictor variables is valuable for variable selection for analysis and explaining the fitted forest (Cutler et al., 2011). While it is standard in numerous applications to run a principal component analysis (PCA) to lessen dimensionality before fitting a classifier, it is conceivable that the principal component analysis capture the important information for prediction issue. For this situation, it might be desirable over acquiring variable significance straight from the algorithm and after that re-fit utilizing just the most vital predictors. Random Forests utilize an uncommon but instinctive measure of variable significance and importance. To do this measurement of the significance of variable k , the algorithm denoted as algorithm 3.6 is implemented for each tree.

Algorithm 3.6 Permutation Variable Importance

To find the importance of variable k, for k=1 to p:

1. (Find $\hat{y}_{i,j}$) For i= 1 to N:

a. Let $\mathcal{g}_i = \{j: (x_i, y_i) \notin \mathcal{D}_j\}$ and Let j_i be the cardinality of \mathcal{g}_i (Algorithm 2). b. Let $\hat{y}_{i,j} = \hat{h}_j(x_i)$ for all $j \in \mathcal{g}_i$

2. (Find $\hat{y}_{i,j}^*$) For j= 1 to J:

a. Let \mathcal{D}_j be the jth bootstrap sample (Algorithm 2).

b. Let $\mathcal{g}_i = \{i: (x_i, y_i) \notin \mathcal{D}_j\}$.

c. Randomly permute the value of variable k for the data points $\{x_i: i \in \mathcal{g}_i\}$ to give \mathcal{p}_j

$\{x_i^*: i \in \mathcal{g}_i\}$

d. Let $\hat{y}_{i,j}^* = \hat{h}_j(x_i^*)$ for all $i \in \mathcal{g}_i$.

3. For i=1 to N:

- For classification:

$$\text{Imp}_i = \frac{1}{j_i} \sum_{j \in \mathcal{g}_i} I(y_i \neq \hat{y}_{i,j}^*) - \frac{1}{j_i} \sum_{j \in \mathcal{g}} I(y_i \neq \hat{y}_{i,j}).$$

- For regression:

$$\frac{1}{j_i} \sum_{j \in \mathcal{g}_i} ((y_i - \hat{y}_{i,j}^*)^2) - \frac{1}{j_i} \sum_{j \in \mathcal{g}} ((y_i - \hat{y}_{i,j})^2).$$

$$\text{Imp} = \sum I(y_i \neq \hat{y}_{i,j}^*)^2.$$

3.4.3 The Artificial Neural Network

This thesis only considers applying the algorithms, specifically the back propagation algorithm.

The pseudocode for the MLP activity proposed by (Koivo, 2008) is as follows:

1. The structure of the network is first defined. In the network, activation functions are chosen and the network parameters, weights and biases, are initialized.
2. The parameters associated with the training algorithm like error goals, maximum number of epochs (iterations), etc., are defined.
3. The training algorithm is called.

4. After the Neural Network has been determined, the result is first tested by simulating the

5. Final validation must be carried out with independent data.

Now let's consider a network with M layers $m=1, 2, \dots, M$

- V_i^m from the output of the i th unit of the m th layer
- V_i^o is a synonym for x_i of the i th input and subscript m layers m 's layers, not patterns
- W_{ij}^m mean connection from V_j^{m-1} to V_i^m .

Algorithm 3.7 ANN Back Propagation

1. Initialize the weights to small random values
 2. Choose a pattern x^d_k and apply to the input layer $V^o_k = x^d_k$ for all k
 3. Propagate the signal through the network $V^m_i = f(\text{net}^m_i) = f(\sum_j W^m_{ij} V^{m-1}_j)$
 4. Compute the deltas for the output layer, $\delta^m_i = f'(\text{net}^m_i)(t^d_i - V^m_i)$
 5. Compute the deltas for the preceding layer for $m=M, M-1, \dots, 2$,
$$\delta^{m-1}_j = f'(\text{net}^{m-1}_j) \sum_i W^m_{ij} \delta^m_i$$
 6. Update all connections
$$\Delta W^m_{ij} = \eta \delta^m_i V^{m-1}_j \quad \Delta W^{\text{new}}_{ij} = W^{\text{new}}_{ij} + \Delta W^m_{ij}$$
 7. Move to step 2 and repeat for the next pattern
-

The feed forward network is the easiest and common type of network (Boritz & Kennedy, 1995 and Kotsiantis, 2007). Training the NN is the method of setting the best weights on the inputs output of the Neural Network with the measured input data. This is compared to the measured

outputs. of each of the units. Back propagation happens to be the most common technique for

calculating the error gradient for a feed forward network (Kotsiantis, 2007). Neural Networks

perform well in applications when the functional form is nonlinear. They are useful in situations

where normal mathematical computation and prior knowledge on the relationship between

inputs and outputs are anonymous. Moreover, an initial step of feature selection before learning

is needed. Multilayer Perceptron (MLP) is applied to predict the impact of IT investment on

banks performance using the study data sets. In this study, linear combination functions and sigmoid transfer functions were used. The S-shaped or Binary sigmoidal function is by far the most common transfer function (Geoffrey & Yau, 2007). The formula for the sigmoid is given:

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (50)$$

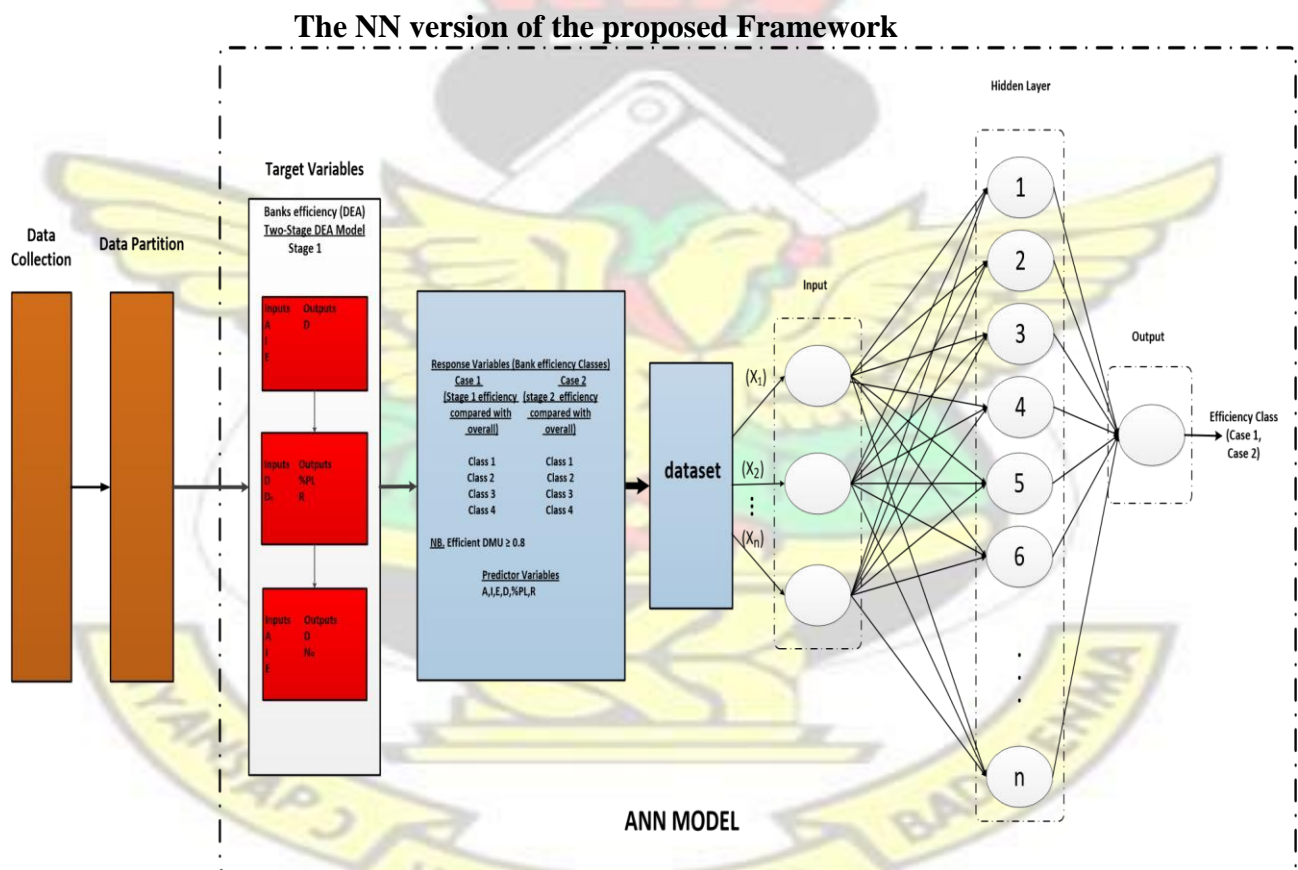


Figure 3.6: The proposed framework (Author's construct)

3.6.3.3 Programming Environment and Dataset

In this part of the methodology, the study presents the data that were used and their descriptions.

The entire dataset was obtained from the dataset that was used in the previous two sections to

build both the DT and RF models. The codes for the NN model building was written in R using the RMiner studio with the “neuralnet” package (Fritsch et al., 2016).

The Output of the NN model

For banks’ performance or efficiency indicators, the study considered two cases (Case 2 and Case 4). Each output consists of four different classes (Classes 1, 2, 3 and 4).

NN Model Inputs

The study also adopted the predictor variables in both cases (Case 2 and Case 4) that were used in the DT and RF model as the NN model inputs.

Number of Hidden Layer and Neurons for the NN

This current study adopted the use of only one hidden layer with five (5) Neurons in a three (3) layer network. This number of hidden neurons was chosen based on the equation $N_h = N_i + 1$ proposed by Tamura & Tateishi (1997) and cited by Vujičić et al. (2016). Where N_h is the number of hidden neurons to be used and N_i is also the number of input neurons. For this study, the number of inputs were 6, meaning $N_i = 6$.

3.4.4 The Logistic Regression Algorithm

Logistic regression algorithm which is the second topmost Machine Learning Algorithm was also used (Appiahene & Missah, 2019). Logistic regression is named for the function used at the core of the method, the logistic function. It measures the correlation between the dependent variable and the one or more independent variables, by estimating probabilities using its core logistic function. These probabilities must then be converted into binary values in order to actually make a prediction. This is the responsibility of the sigmoid function which is an Sshaped curve that can take any real-valued number and map it into a value between the range of 0 and 1, but never exactly at those limits. This value between 0 and 1 will then be transformed into either 0 or 1 using a threshold classifier. The figure below illustrates the steps that logistic regression goes through to give you the desired output (efficiency class).

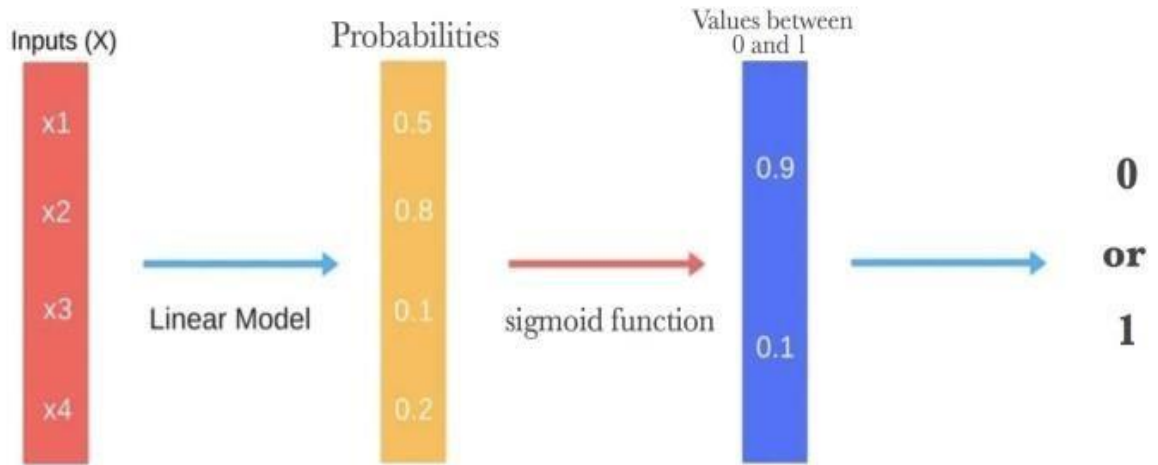


Figure 3.7 : The Logistic Regression Steps (Donges, 2019)

It is important to understand that the goal of an analysis using logistic regression is the same as that of any model-building technique used in Machine Learning: To find the best fit and most parsimonious. What actually makes a logistic regression algorithm different from the first topmost Machine Learning Algorithm, Linear Regression Algorithm is the outcome variable. In the logistic regression model, the outcome variable is binary or dichotomous. Logistic regressions work with **odds** rather than proportions. The odds are simply the ratio of the proportions for the two possible outcomes. The formula for the logistic regression is given by:

$$\log_e(1 - p_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i, i = 1, \dots, n \quad (51)$$

Where β s are the regression parameters, the Xs are the explanatory variables and n is the sample size.

The Logistic Regression Algorithm (Angel et al., 2016) adapted for this study is described below. This algorithm was implemented in R codes using the package “elrm” Zamar et al. (2007) .

Input:

1. Training algorithm L(logistic regression)
2. Sample matrix X
3. Labels vector $y=[1,\dots,K]$
4. Initial regressor parameters vectors θ_i

For $i=1:K$

 Create a new binary vector y_i for each label where $y_i=1$ if it belong to the label and $y_i = 0$ if it does not belong.

Apply L to X to find θ_i

Output:

θ_i Parameters vector for each regressor.

To implement the Logistic Regression Algorithm in R, the study adopted the following steps:

- Data Pre-processing step
- Fitting Logistic Regression to the Training set
- Predicting the test result □ Test accuracy of the result
- Visualizing the test set result.

3.4.5 Metrics for Evaluating Machine Learning Algorithms

The comparison of machine learning algorithms models should normally be done using the standard machine learning evaluation metrics and some of the evaluation metrics (Yao et al., 2017) that are usually used include but not limited , the root mean squared error (RMSE), mean (rBIAS), relative mean separation (rMSEP), mean absolute percentage error (MAPE), and root

mean squared prediction error (RMSPE). These standard machine learning evaluation metrics

are defined as:

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{z}_i - z_i)^2} \\ MAE &= \frac{1}{m} \sum_{i=1}^m |\hat{z}_i - z_i| \end{aligned} \tag{52}$$

$$(53) rBIAS = \frac{1}{m} \sum_{i=1}^m (\hat{z}_i - z_i)$$

mz

$$rMSEP = \sum_{i=1}^m (z_i - \hat{z}_i)^2 / \sum_{i=1}^m (z_p - z_i)^2 \quad (55)$$

$$MAPE = \left(\frac{1}{m} \sum_{i=1}^m \left| \frac{\hat{z}_i - z_i}{z_i} \right| \right) 100 \quad (56)$$

$$RMSPE = \sqrt{\frac{\sum_{i=1}^m (\hat{z}_i - z_i)^2}{m}} \quad (57)$$

Where, m is the total number of observations we want to validate, z_i is the data indexed by i , \hat{z}_i is the prediction value, \bar{z} and \bar{z}_p are the arithmetic mean of the observations and predictions respectively.

(54)

CHAPTER 4

RESULTS AND DISCUSSIONS

4.0 Introduction

This chapter present and discusses results on the following: the application Decision Tree Algorithms, Random Forest Algorithms and Artificial Neural Network models for predictions; (2) the use of DEA for efficiency measurement and classification of the banks into various classes using the proposed Bank Classification Algorithm (3) the application of the four proposed machine learning models in predicting the efficiency of banks.

4.1 DISCUSSIONS OF THE PREDICTION OF THE BANKS USING DECISION TREE ALGORITHM

Case 2-The Deposit Stage

According to the result, the DT model gave a high prediction accuracy of 50.8% and kappa value of 29.6% with a P-Value of 0.003298. Thus, the DT model was able to predict classify the banks using the classification classes as response variables at an accuracy rate of about 50% in the Case 2. Using the RMSE, MAE, MAPE, RMSPE and rBIAS as the performance evaluation measurement of the model, the following results were suggested: RMSE=0.70181, MAE=0.492537, MAPE=49.2537313, RMSPE=0 and rBIAS=0.492537.

The detailed results of the performance measurement of the DT model in classifying the various classes under Case 2 are shown in the confusion matrix generated by the Case 2 analysis below:

Confusion Matrix and Statistics

		Reference			
Prediction	Class	Class 1	Class 2	Class 3	Class 4
1	2	1	3	4	
Class 2		9	25	8	17
Class 3		3	2	10	0
4	7	10	2	31	Overall
Statistics					

Accuracy : 0.5075 95%
 CI : (0.4198, 0.5948)
 No Information Rate : 0.3881
 P-Value [Acc > NIR] : 0.003298

Kappa : 0.2958

Case 4 –The Investment Stage

Under the Case 4, the DT model had a very low accuracy of 38.8% and kappa value of 16.6% with a P-Value of 0.08513 as compared to Case 2. In this the case of the DT in case the model cannot be used since its P-Value is above the threshold for significance.

Using the RMSE, MAE, MAPE, RMSPE and rBIAS as the performance evaluation measurement of the model, the following results were suggested: RMSE=0.782266, MAE=0.61194, MAPE=61.19403, RMSPE=0 and rBIAS=0.61194. The detailed results of the performance measurement of the DEA-DT model in predicting the various classes under

Case 4 are shown confusion matrix below:

Confusion Matrix and Statistics

		Reference				
Prediction	Class 1	Class 2	Class 3	Class 4	Class 5	
1	2	4	4	7		
Class 2		3	7	2	4	
Class 3	10	11	15	5	Class	
4	14	12	6	28		

Overall Statistics

Accuracy : 0.3881 95%
 CI : (0.3052, 0.476)
 No Information Rate : 0.3284 P-Value
 [Acc > NIR]: 0.08513
 Kappa : 0.1658
 McNemar's Test P-Value: 0.01757

4.2 DISCUSSIONS OF THE PREDICTION OF THE BANKS USING RANDOM FOREST ALGORITHM

Case 2-The Deposit Stage

The out-of-bag estimates of the error rate given by $E_{oob} = \frac{1}{N} \sum_{i=1}^N I_N(y_i \neq \mathcal{F}_{oob}(x_i))$ were used to choose the optimum Random Forest parameters (formula = Case2 ~ data = TrainSet, ntree = 500, mtry = 2, importance = TRUE). According to the result, the RF model gave a high overall prediction accuracy of 58.2%. This means that the Random Forest model was able to predict the efficiency of the new 30% bank branches at an overall accuracy rate of about 58% in the Case 2. Using the RMSE, MAE, MAPE, RMSPE and rBIAS as the performance evaluation measurement of the model, the following results were suggested: RMSE=0.64646, MAE=0.41791, MAPE=41.79104, RMSPE=0 and rBIAS=0.41791. The following is also the confusion matrix generated by Case 2 analysis:

Confusion Matrix and Statistics

		Reference			
Prediction	Class	Class 1	Class 2	Class 3	Class 4
1	2	1	5	1	
Class 2		7	16	4	14
Class 3		6	2	16	1
Class 4		3	8	4	44

Overall Statistics

Accuracy : 0.5821

95% CI : (0.4938, 0.6667)

No Information Rate : 0.4478

P-Value [Acc > NIR] : 0.001219

Kappa : 0.3959

Mcnemar's Test P-Value : 0.138147

Case 4-The Investment Stage

For Case 4 out-of-bag estimates of the error rate(E_{oob}) were used to select the optimum Random Forest parameters (formula = Case4 ~., data = TrainSet, ntree = 500, mtry = 6,

importance = TRUE). Under the Case 4, the RF model had a lower prediction accuracy of 40.3% as compared to the Case 2 which had 58.21%. Using the RMSE, MAE, MAPE, RMSPE and rBIAS as the performance evaluation measurement of the model, the following results were suggested: RMSE=0.772667, MAE=0597015, MAPE=59.7015, RMSPE=0 and rBIAS=0597015. Confusion Matrix and Statistics

	Reference			
Prediction	Class 1	Class 2	Class 3	Class 4
Class 1	0	3	3	2
Class 2	4	4	6	6
Class 3	6	8	17	6
Class 4	21	8	7	33
Overall Statistics				

Accuracy : 0.403
 95% CI : (0.3192, 0.4911)
 No Information Rate : 0.3507
 P-Value [Acc > NIR] : 0.120256
 Kappa : 0.1615 Mcnemar's Test P-
 Value : 0.007651

4.3 DISCUSSIONS OF THE PREDICTION OF THE BANKS USING ARTIFICIAL NEURAL NETWORK

Case 2-The Deposit Stage

When the NN model which was trained and validated using 310 (70%) bank branches dataset was used on the other 134 bank branches, the NN model was able to classify the banks into their respective classes at an accuracy rate of 14.2%. Using the RMSE, MAE, MAPE, RMSPE and rBIAS as the performance evaluation measurement of the model, the following results were suggested: RMSE=1.176902, MAE=1.0421235, MAPE=57.45151114, RMSPE=1.179884 and rBIAS=-0.01943.

The following is also the confusion matrix of the NN analysis:

Confusion Matrix and Statistics

	Reference			
Prediction	1	2	3	4
1	0	0	0	0
2	0	0	0	0
3	24	27	19	64
4	0	0	0	0

Overall Statistics

Accuracy : 0.141791 95% CI :
 (0.0875778, 0.2125375)

No Information Rate : 0.4776119

P-Value [Acc > NIR] : 1

Kappa : 0

Case 4-The Investment Stage

Under Case 4, the NN model had an accuracy of 16.42% as compared to Case 2 which was a little higher. Using the RMSE, MAE, MAPE, RMSPE and rBIAS as the performance evaluation measurement of the model, the following results were suggested:

RMSE=1.241123, MAE=1.11657, MAPE=65.49859, RMSPE=1.227733 and rBIAS=-0.02396. The following is also the confusion matrix of the NN analysis:

Confusion Matrix and Statistics

		Reference			
Prediction		1	2	3	4
1		0	0	0	0
2		0	0	1	2
3		33	23	22	53
4		0	0	0	0

Overall Statistics

Accuracy : 0.1641791

95% CI : (0.1058435, 0.2379513)

No Information Rate : 0.4104478

P-Value [Acc > NIR] : 1

Kappa : -0.009009

4.4 DISCUSSIONS OF THE BANKS' EFFICIENCY SCORES AND CLASSIFICATION

4.2.1 Bank Efficiency Determination Using the Adapted DEA Two-Phase BuildHull

Algorithm

For each Bank, the technical efficiencies in both stages (Deposit Stage, Investment Stage) and their corresponding overall efficiencies were analyzed using CCR DEA algorithm proposed by Mehrabiana (2013) and depended on non-Archimedean Charnes-Cooper-Rhodes (CCR). Using the 444 bank branches, the efficiency of each bank branch at each stage was analyzed using the scenarios as discussed one after the other below.

Scenario 1: When an efficient unit is defined as a unit with an efficiency score of 1 unit or 100%.

For banks efficiency in terms of utilizing their resources to collect deposit from customers , only 14(3.15%) bank branches were efficient (had 100% efficiency). Just 33(7.43%) bank branches had an efficiency scores of between 80% to 99%,1 (0.23%) bank branch also had efficiency score of between 70 to 79, 21(4.73%) had efficiency score of between 60 to 69; 19 (4.28%) had between 50 to 59 and 356 (80.18%) had an efficiency score below 50%. This 356(80.18%) number of bank branches confirms the fact that a lot of Ghanaian banks are not efficient in using their resources to collect deposits from customers. This confirms the fact that most banks in Ghana are struggling to meet the minimum capital requirements set by the central bank (Bank of Ghana) in 2017 (Addison, 2017, 2018a and 2018b) .The results of the deposit stage efficiency is shown in Figure 4.1.

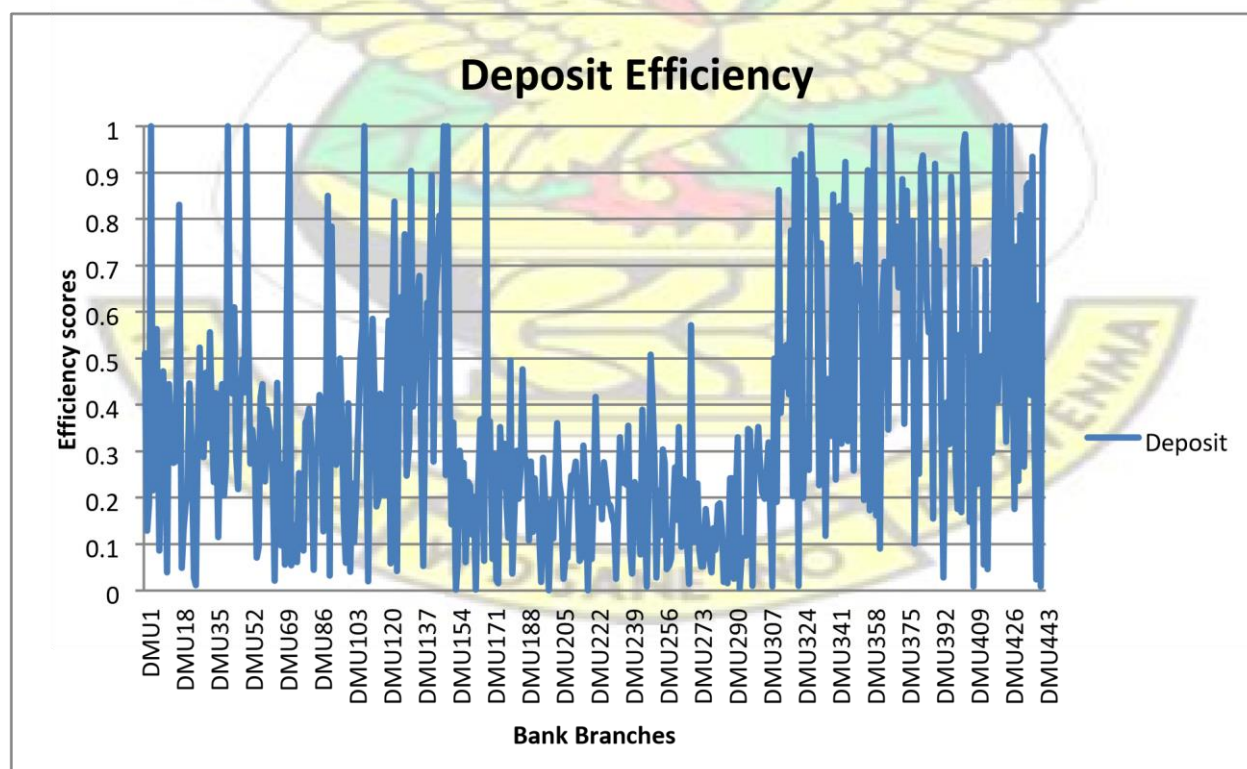


Figure 4.1: A graph showing the deposit efficiency scores of the 444 DMUs (Author's construct)

For banks efficiency with regarding investing customers' deposits shown in Figure 4.2, only

1bank (DMU200) was efficient in investing the deposit to generate profit for the banks while only 1 bank (DMU219) had an efficiency score of between 80% to 100%. This result suggests that close to 99.5% of Ghanaian bank branches that were considered for the study were not efficient in investing customers deposits. This also confirms reports of crises that have hit the Ghanaian banking industry with their issues such as an alleged managers and board of directors squandering depositors' money without investing them Abubakar (2018), and has also lead to about seven (7) universal banks collapsing in 2017 and 2018(Addison, 2018a; Addison, 2018b).

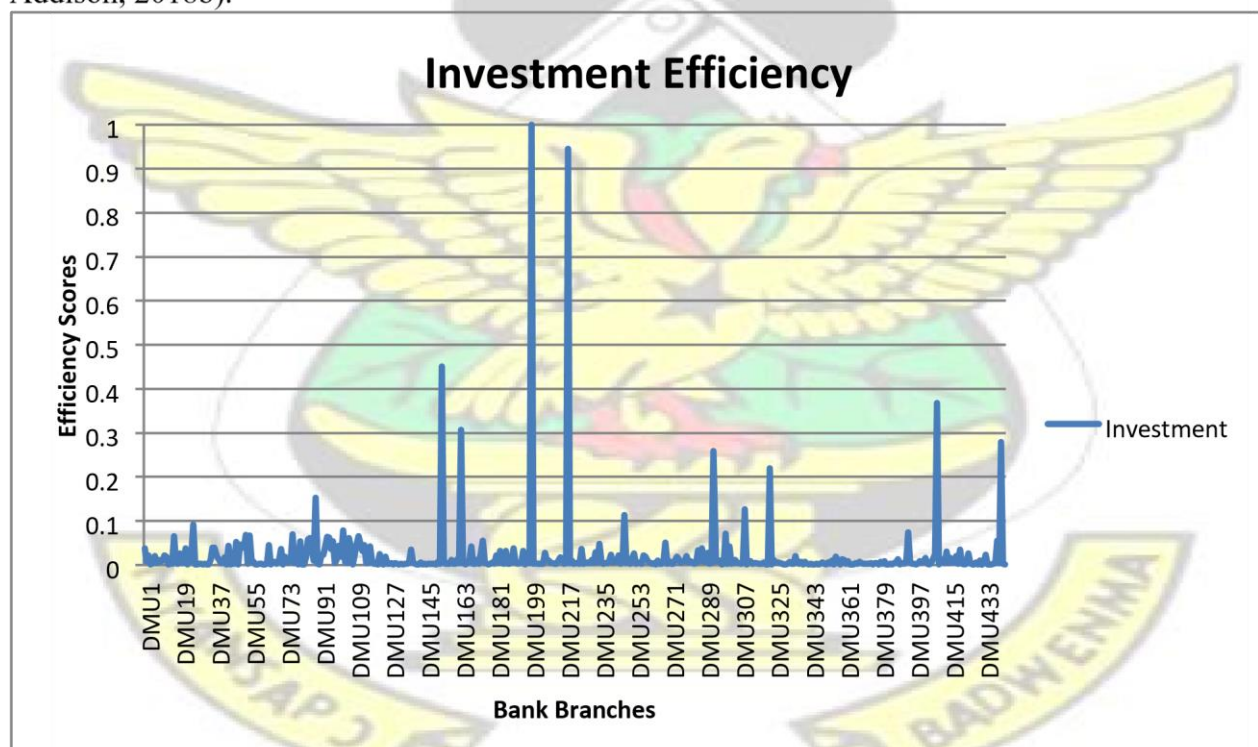


Figure 4.2: A graph showing the investment efficiency scores of the 444 DMUs (Author's construct).

For Overall efficiency in the entire banking operations also shown in Figure 4.3, 79(17.79%) bank branches were efficient (had 100% efficiency score) with the majority (290 representing, 65.32%) of them having an efficiency score of between 80% to 99%. 4(0.9%) bank branches had efficiency score of between 70 to 79%, 32(7.21%) had an efficiency scores of between 60 and 69% and finally 39(8.78%) branches had between 50 to 59%. In terms of overall

efficiency, there was no bank branch that had less than 50% efficiency score evident in the

Figure 4.3.

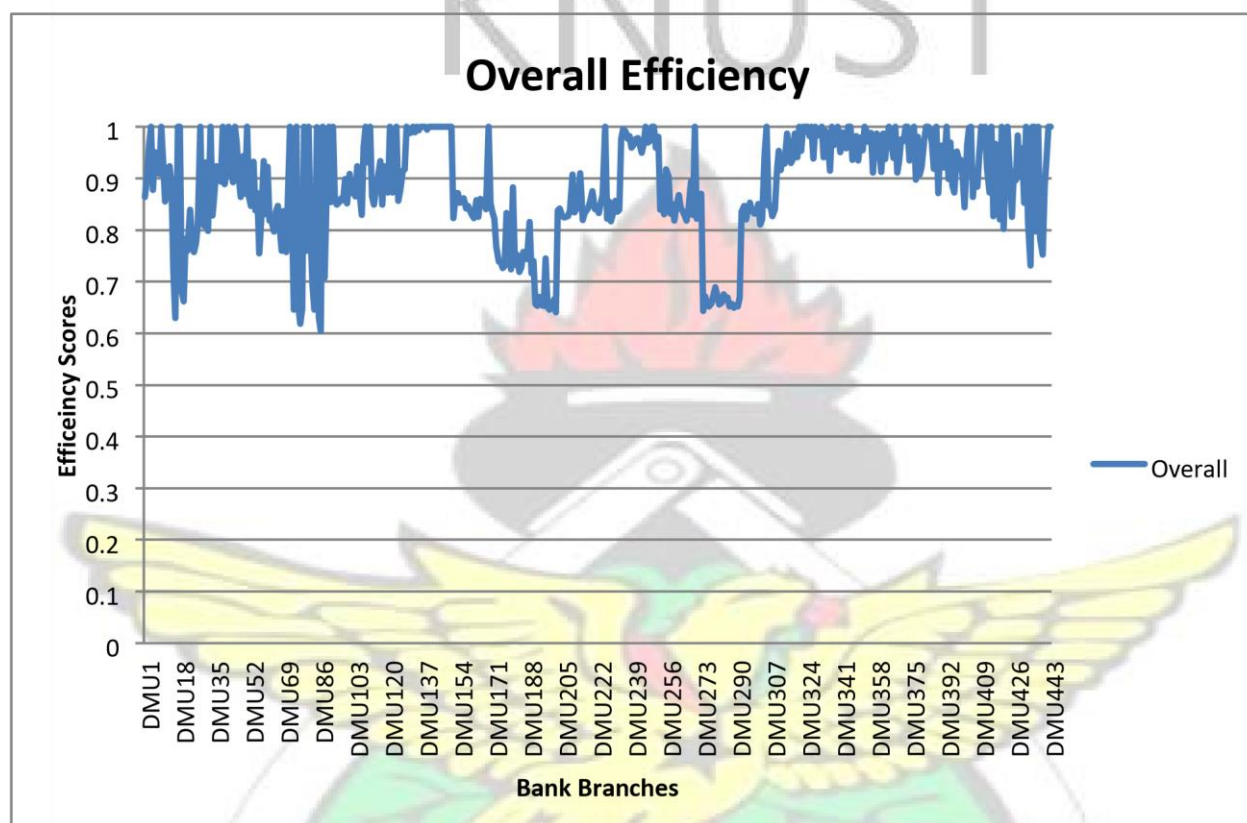


Figure 4.3: A graph showing the overall efficiency scores of the 444 DMUs (Author's construct)

This analysis means that even though most bank branches in Ghana do not experience a higher percentage of efficiency in collecting deposit and investing the deposit, they still enjoy higher overall efficiency. The results suggest that banks in Ghana should identify ways of improving their efficiencies in both the deposit stage and investment stage and should not only rely on their overall efficiency scores as means of measuring their performance and success.

In real a life situation, it is very difficult to have units or departments attaining 100% efficiency and bank branches in Ghana are no exception. Normally, in grading of Systems in terms of their efficiency, systems with efficiency value between 80% and 100% are also considered to be excellent (efficient) in their performance. This is also evident in the Ghanaian banking sector

where the central bank (Bank of Ghana) always has minimum capital requirement for banks and other financial institutions (Addison, 2017, 2018a, 2018b).

This means that there is always a “cutoff point” for banks to meet in order to be efficient and remain competitive in the banking sector. Based on this, the author also inferred and considered banks with an efficient value of 80% or more as efficient. The study therefore adopted the efficiency “cutoff point” ($\text{DMU efficiency} \geq 0.8$) suggested by Wu (2006) and cited by Chen (2016) and Santos et al. (2017) to discuss the efficiency scores in all the stages as follows: **Scenario 2: When an efficient unit is defined as a unit with an efficiency score of 0.8 or more (80-100%).**

Under this scenario, 47 (10.59%) banks were efficient in collecting deposit from customers while only 2 banks (bank with DMU200 and bank with DMU219) were efficient in investing the collected deposit to generate profit. A good number of bank branches 369 (83.11%) were efficient at the entire banking operation (overall stage).

In all, the analysis and discussion of the results, a certain trend was emerging as shown in Figure 4.4. Thus, the banks with higher deposit efficiency were likely to have a higher overall efficiency and that deposit efficiency was contributing to the overall efficiency of a bank as shown in Figure 4.4.

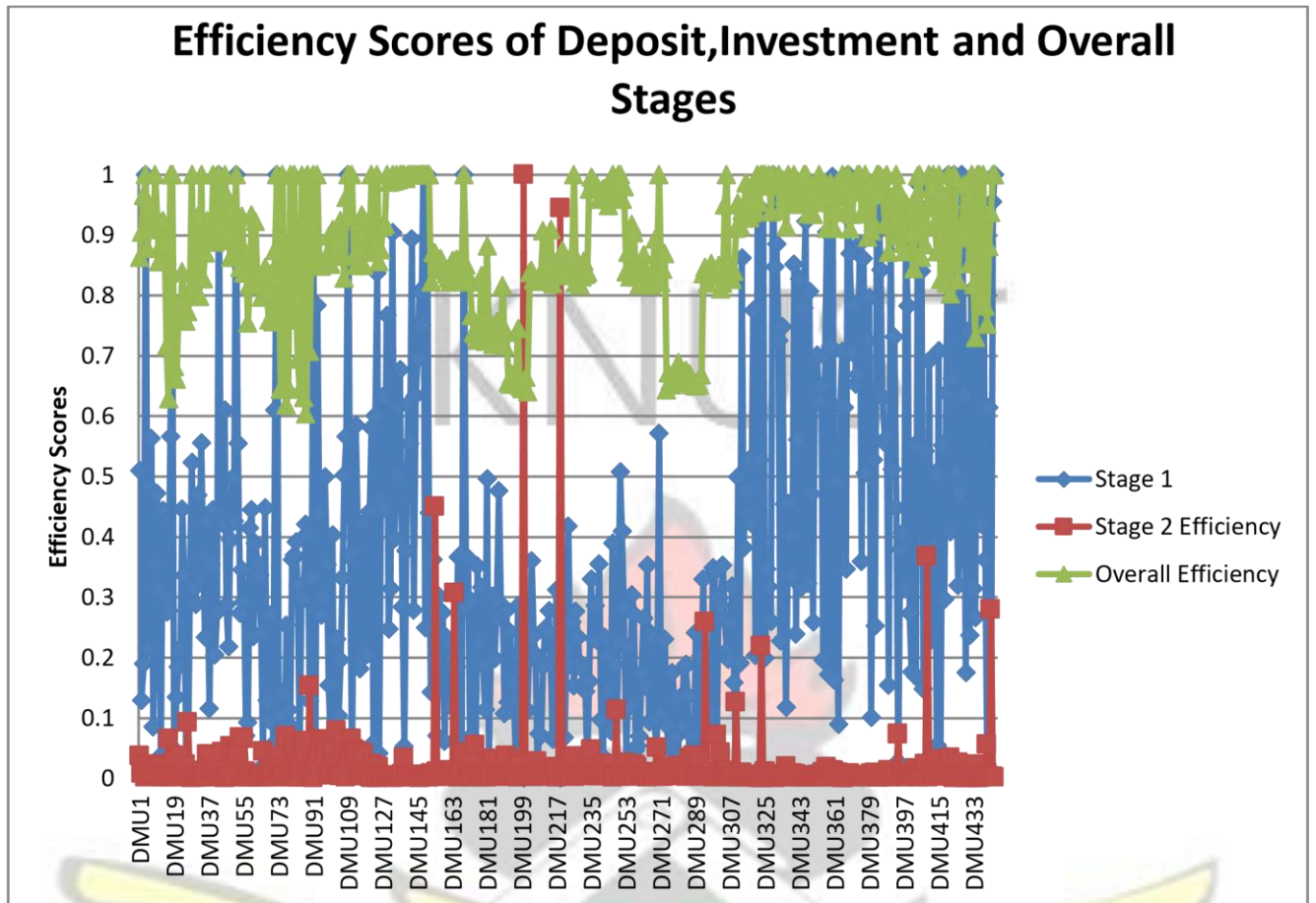


Figure 4.4: A graph showing the efficiency scores of deposit, investment and overall stages of the 444 DMUs (Author's construct)

This is a graph representing the 444 banks used in the study. From Figure 4.4, banks such as DMU4, DMU10, DMU18 etc. had the highest deposit efficiencies and still had the highest overall efficiencies. What is also evident from the graph is that a lot of the banks were not efficient in investing deposit collected from customers, except bank DMU200 and bank DMU219, also had efficiency of 100% and 90% respectively.

4.2.2 Using the proposed Banks' Classification Algorithm (BC Algorithm) to classify Banks based on their Efficiencies Scores.

The various banks (DMUs) were classified into classes using the BC Algorithm proposed by this study using both the deposit stage efficiency and investment stage efficiencies. In this first

instance (Case 1 and Case 3), efficient bank was defined using DEA score as 100% (1 unit). As suggested earlier, it is very difficult to have units or departments attaining 100% efficiency in real life situation and bank branches in Ghana are no exception. The study, therefore, adopted the efficiency “cutoff point” (DMU efficiency ≥ 0.8) suggested by Wu (2006) and cited by Chen (2016) and Santos et al. (2017) by classifying the bank branches based on their efficiency scores in the two stages using this “cutoff point”. Using the overall efficiency as the basis, the efficiency of both stage I (banks efficiency in collecting deposit from customers) and the stage II (banks efficiency in investing their deposit into securities and given out loans) were categorized into classes Wang et al. (1997) and without any order. In this case four scenarios were used and the results are shown in Appendix B.

Case 1: When stage I efficiency was categorized on the basis of overall efficiency and in both cases efficient unit is one with a 100% (1) efficiency value. The classifications are as follows:

- Class 1: Banks efficient in collecting deposit but were not able to attain overall efficiency.
- Class 2: Banks that realized overall efficiency even though they were not efficient in collecting deposit.
- Class 3: Banks that were efficient in collecting deposit and still had overall efficiency.
- Class 4: Banks inefficient in collecting deposit and cannot also achieve overall efficiency.

Case 2: When stage I efficiency was categorized on the basis of overall efficiency and in both cases efficient unit is one with 80%-100% (0.8-1) efficiency value.

The classifications are below:

- Class 1: Banks efficient in collecting deposit but were not able to attain overall efficiency.
- Class 2: Banks that realized overall efficiency even though they were not efficient in collecting deposit.
- Class 3: Banks that were efficient collecting deposit and still had overall efficiency.
- Class 4: Banks inefficient in collecting deposit and cannot also achieve overall efficiency.

Case 3: When stage II efficiency was categorized on the basis of overall efficiency and in

both cases efficient unit is one with 100% (1) efficiency value. The classifications are as follows:

- Class 1: Banks efficient in investing deposit but were not able to attain overall efficiency.
- Class 2: Banks that realized overall efficiency even though they were not efficient in investing deposit.
- Class 3: Banks that were efficient in investing deposit and still had overall efficiency.
- Class 4: Banks inefficient in investing deposit and cannot also achieve overall efficiency.

Case 4: When stage II efficiency was categorized on the basis of overall efficiency and in both cases efficient unit is one with 80%-100% (0.8-1) efficiency value. The classifications are as follows:

- Class 1: Banks efficient in investing deposit but were not able to attain overall efficiency.
- Class 2: Banks that realized overall efficiency even though they were not efficient in investing deposit.
- Class 3: Banks that were efficient in investing deposit and still had overall efficiency.
- Class 4: Banks inefficient in investing deposit and cannot also achieve overall efficiency.

Using the proposed algorithm for the classification, the following were discovered:

4.2.2.1 For Cases 1 and 2: When the efficient value was 100%(1) ,only 1 DMU (DMU 427) was in class 1 while the majority of the DMUs representing 82% were in class 4, followed by class 2 and class 3. This means that the majority (364) of the commercial bank's branches (DMU's) were not 100% efficient in collection deposit from customers and also in their overall efficiency. Only just about 13 DMUs were efficient in the collection of deposit from customers. The entire results of this case 1 analysis are shown in Table 4.1 with its corresponding bar chart also shown in Figure 4.5.

When the cutoff for efficient DMU was drop from 100% to 80-100%, the majority (72.5%) of the DMUs were now in class 2 which means that the majority of the DMUs even though were not efficient in collecting deposits (stage I), they were able to achieve overall efficiency. Just

about 10.6% of the DMUs were efficient in utilizing resources for collecting deposit from customers and also using the resources to achieve overall efficiency. The results of this are also presented in Table 4.2 and Figure 4.6.

Table 4.1.: Efficiency Classes (case 1) (Authors construct).

	Frequency	Percent	Valid Percent	Cumulative Percent
Class 4	364	82.0	82.0	82.0
Class 1	1	.2	.2	82.2
Class 2	66	14.9	14.9	97.1
Class 3	13	2.9	2.9	100.0
Total	444	100.0	100.0	

Bar Graph of Efficiency Classes (Case 1)

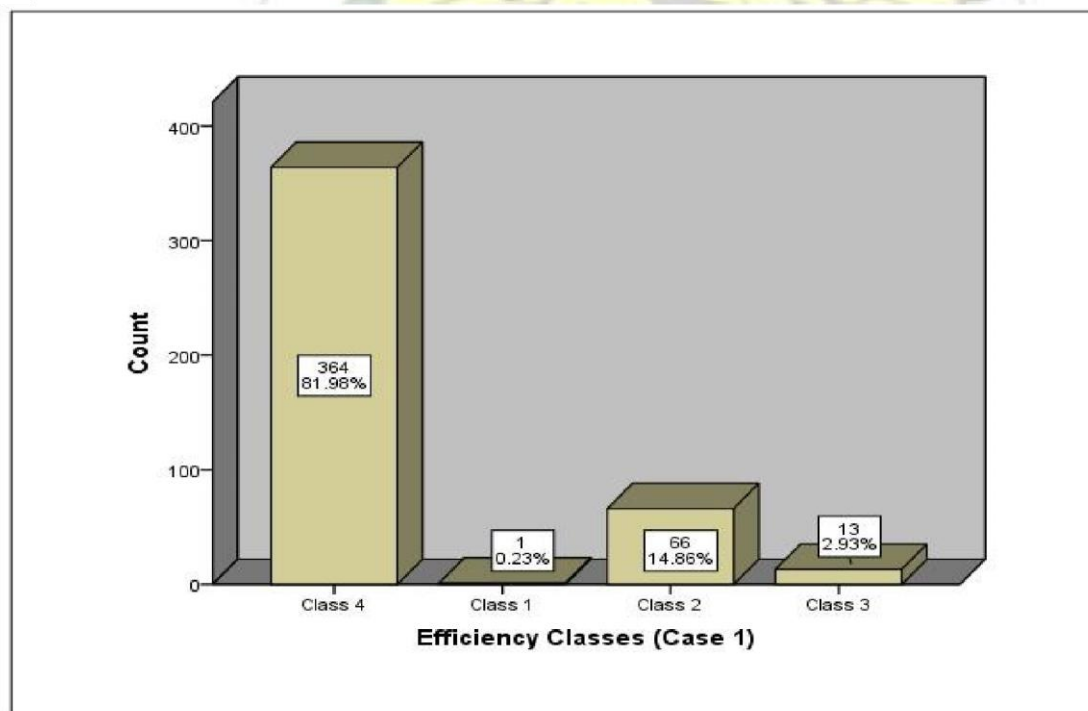


Figure 4.5: A bar graph showing the proposed efficiency classes (Case 1) and the number of DMUs (Authors construct).

Table 4.2: Efficiency classes (Case 2) (Authors construct).

Efficiency classes (Case 2)				
	Frequency	Percent	Valid Percent	Cumulative Percent
Class 4	75	16.9	16.9	16.9
Class 2	322	72.5	72.5	89.4
Class 3	47	10.6	10.6	100.0
Total	444	100.0	100.0	

Bar Graph of Efficiency Classes (Case2)

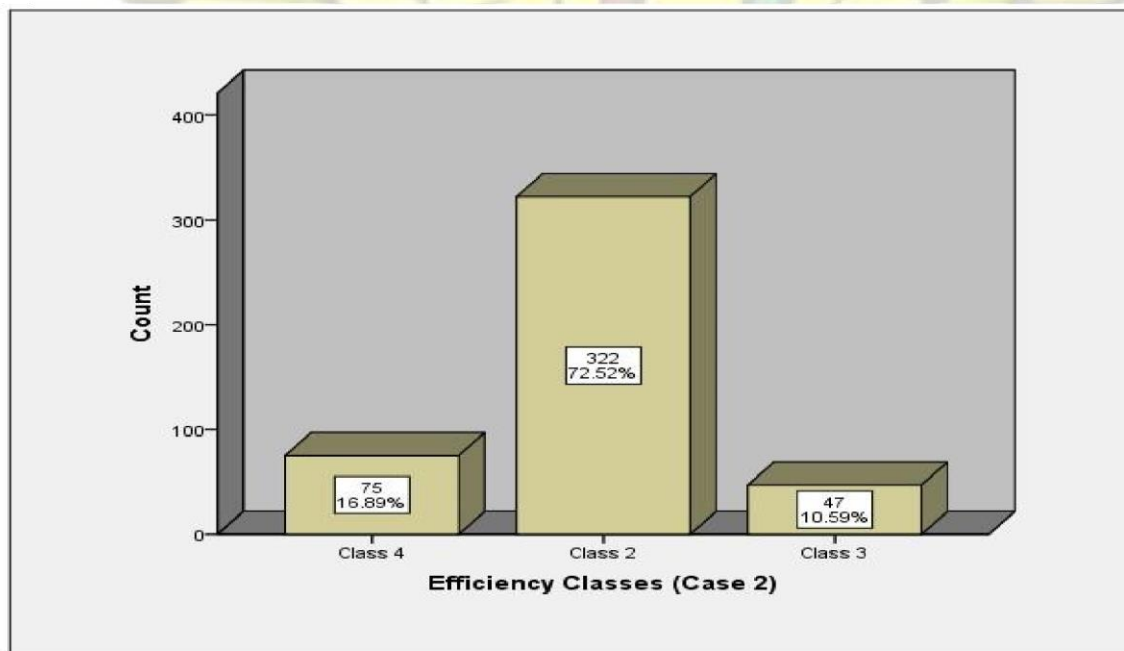


Figure 4.6: A bar graph showing the proposed efficiency classes (Case 2) and the number of DMUs (Authors construct).

For Cases 3 and 4: When the efficient DMU value was at 100% (1), only 1 DMU (DMU 201) was in class 1 while the majority of the DMUs representing 82% were not 100% efficient in investing deposit from customers and giving out loans and also in their overall efficiency

followed by class 2 and there was no DMU in class 3. Thus no DMU was 100% efficient in investing their deposit. 79 DMUs (17.8%) were able to achieve overall efficiency even though there were not efficient in investing their deposits.

The entire results of this case 3 analysis are shown in Table 4.3 below with its corresponding bar chart also shown in Figure 4.7. When the cutoff for efficient DMU was drop from 100% to 80-100%, majority 368 (82.9%) of the DMUs were now in class 2 which means that the majority of the DMUs even though were not efficient in using their resources to invest their deposit (stage II), they were able to achieve overall efficiency. Just about 16.7% (74) of the DMUs were not efficient in investing their deposit from customers and also using resources to achieve overall efficiency. The results of this are also shown in Table 4.4 and Figure 4.8.

Table 4.3: Efficiency classes (Case 3) (Authors construct).

Efficiency Classes (Case3)					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Class 4	364	82.0	82.0	82.0
	Class 1	1	.2	.2	82.2
	Class 2	79	17.8	17.8	100.0
	Total	444	100.0	100.0	

Bar Graph of Efficiency Classes (Case 3)

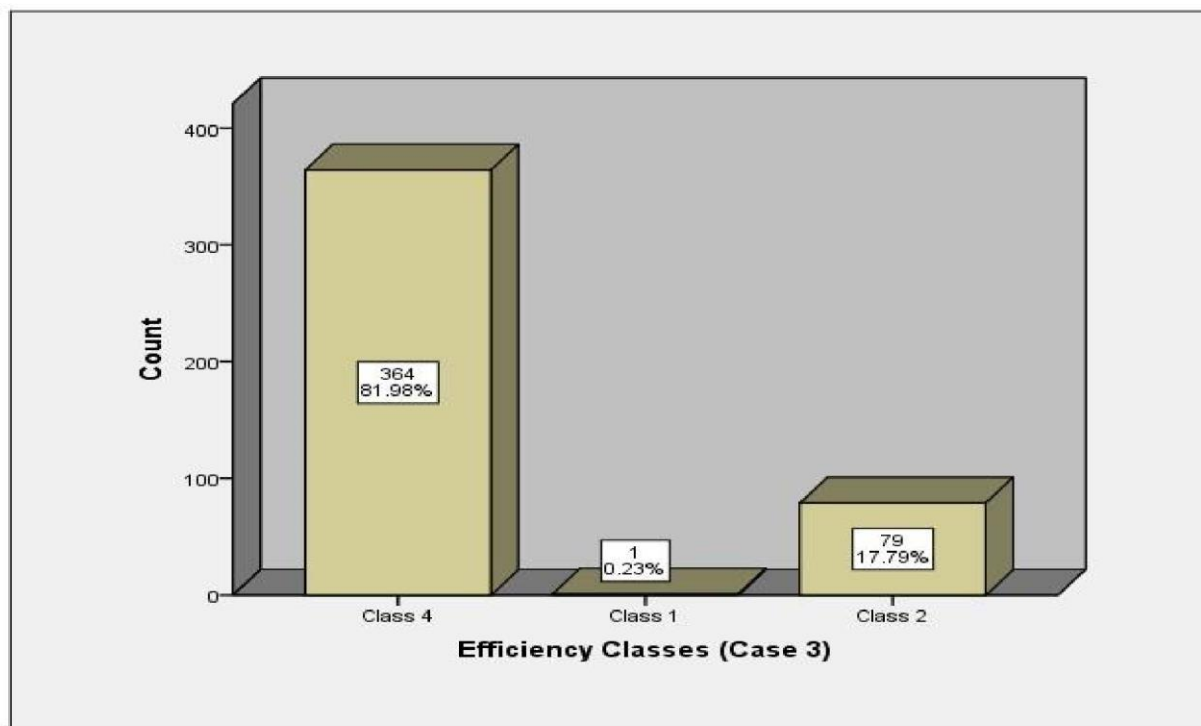


Figure 4.7: A bar graph showing the proposed efficiency classes (Case 3) and the number of DMUs. (Authors construct).

Table 4.4: Efficiency classes (Case 4) (Authors construct).

	Frequency	Percent	Valid Percent	Cumulative Percent
Class 4	74	16.7	16.7	16.7
Class 1	1	.2	.2	16.9
Class 2	368	82.9	82.9	99.8
Class 3	1	.2	.2	100.0
Total	444	100.0	100.0	

Bar Graph of Efficiency Classes (Case 4)

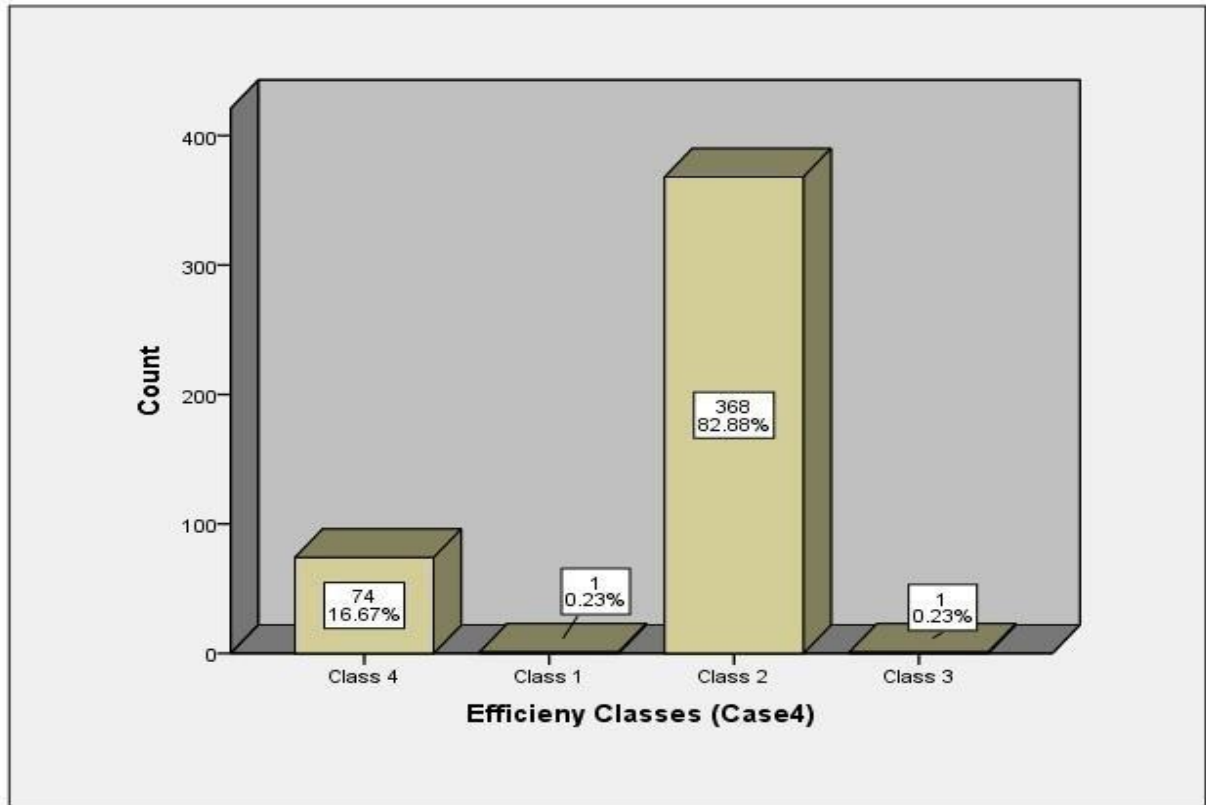


Figure 4.8: A bar graph showing the proposed efficiency classes (Case 4) and the number of DMUs (Authors construct).

4.5 DISCUSSION OF THE PROPOSED BANK EFFICIENCY PREDICTION FRAMEWORK

The proposed framework that was developed for building the predictive model was done under two scenarios. The first scenario was when the model was developed taking into consideration Decision Tree and Random Forest Algorithms. In both Decision Tree and Random Forest Algorithms, the dataset used for building the predictive models consisted of predictor variables and target or response variables. The second scenario, the case of building a predictive model using Artificial Neural Network, the predictor variables used for both Decision Tree and Random Forest Algorithms were considered as input while its target or response variables as in the case of Decision Tree and Random Forest were also considered as output for the Neural Network model. The study, therefore, developed the model for building a predictive model under two cases considered above but they were all working using the same concept. The discussion of the proposed framework in this chapter takes into consideration the two scenarios;

thus the model for both Decision Tree/Random Forest Algorithm model development and Artificial Neural network predictive model development.

At the preliminary stage (stage I), raw data were collected from the banks. After data collection, it is preprocessed in stage II before finally entering the stage where the predictive model development takes place. In this case, the financial data such as the Cedi value of IT Expenditure (I), Fixed Asset(A), Total Deposit(D), Profit(R), Rate of Performing Loans(%PL), and finally the number of employees (E) from banks were used to calculate the efficiencies of the various banks (DMUs) using the CRS technology. The efficiencies of the banks at both deposit and investment stages were classified into classes (Classes 1 to 4) without any order but on the bases of their overall efficiency using the proposed Bank Classification Algorithm.

This then becomes the target or dependent variables.

Now the banks financial data such the IT expenditure, fixed Assets etc. were used as predictor variables to predict the efficiency scores (Classes) of each bank branch for building the models.

In the case of the target variables, the efficiencies of the banks, grouped into classes (Classes 1-4) were based on scenarios or cases in the study, hence two cases were used where efficient unit is one with efficiency score greater than or equal to 0.8.

This final dataset for building the model is then randomly divided into two, 70% were train or build and validate the model using K-fold cross-validation while the remaining 30% (test dataset) was used to test the models. Thus in the case of the banks, 70% of the DMUs which were selected randomly from the total dataset were used to build and validate the model is now used to predict the efficiency of the other 30% banks (DMUs). During the model building, the dataset also goes through rule extraction and finally building the classifier. If the suggested model can work with as accepted or high accuracy, the model can be used to estimate the efficiencies of new banks.

4.5.1 DISCUSSIONS OF THE PREDICTION OF THE BANKS EFFICIENCIES USING THE DEA-DT MODEL

A total of 444 bank branches (DMUs) on two separate scenarios, Case2 and Case 4 were implemented in a DEA-DT model using R codes in RMiner studio. In each, there were four dependent variables (Classes 1 to 4) and six (6) independent variables. Therefore, the target variables as performance (efficiency) measure of each category were predicted as either in Class 1, Class 2, Class 3 or Class 4. The performance measure of the DEA- DT model was done by using a confusion matrix (Delen et al., 2013), RMSE, MAE, MAPE, RMSPE and rBIAS. The dataset for this analysis in both cases was divided randomly into 70% (training and validation dataset) and 30% (test dataset). For the performance analysis of the proposed

DEA-DT model, the test dataset was used.

To evaluate the prediction accuracy of the model, K-fold cross validation was implemented to lessen the bias of sampling data and ensuring model error randomness. K-fold cross validation randomly divides data into k subsets and one subset is used as testing data and k-1 subsets are used as training data (Yousaf, 2016) . This process is repeated k times to cover all data. The study used 10-fold cross validation in the analysis. In addition to the k-fold cross validation the study also use repeated cross validation where data partitioning was done to split the dataset into training (70%) and testing (30%). To estimate the models' performance, RMSE, MAE, MAPE, RMSPE and rBIAS were used. Analysis of the prediction by the DEA-

DT model was done in two cases (Cases 2 and 4). In the first case, the Case 2 variables were the dependent variables and the result was compared with the classes designated by the CCR during the DEA stage. The entire results of the DEA-DT prediction of both cases are shown in Appendices C and D.

4.5.1 The Bank branches (132 DMUs) efficiency analysis using the DEA-DT model

Result

4.5.1.1 Case 2 –The Deposit Stage

According to the result, the combined DEA-DT model gave a high prediction accuracy of

91.04% and kappa value of 80.2% with a P-Value of 5.296e-07. This means that the proposed DT model was able to predict the efficiencies of the 30% banks at an accuracy rate of about 91%% in the Case 2. The detailed results of the performance measurement of the DEA-DT model in predicting the various classes under Case 2 are shown in the confusion matrix generated by the Case 2 analysis below:

Confusion Matrix and Statistics

	Reference			
Prediction Class	Class 2	Class 3	Class 4	Class
2	89	2	0	
Class 3	7	9	0	Class
4	3	0	24	

Overall Statistics

Accuracy : 0.9104
 95% CI : (0.8488, 0.9529)
 No Information Rate : 0.7388
 P-Value [Acc > NIR] : 5.296e-07
 Kappa : 0.802

4.5.1.2 Case 4-The Investment Stage

Under the Case 4, the proposed DEA-DT model had the highest prediction accuracy of 97.8% and kappa value of 92.7% with a P-Value of 1.48e-08 as compared to Case 2. The detailed results of the performance measurement of the DEA-DT model in predicting the various classes under Case 4 are shown confusion matrix below:

Confusion Matrix and Statistics

	Reference			
Prediction Class	Class 1	Class 2	Class 3	Class 4
1	0	0	0	0
Class 2	0	107	0	0
Class 3	0	0	0	0
4	0	3	0	24

Overall Statistics

Accuracy : 0.9776
95% CI : (0.936, 0.9954)
No Information Rate : 0.8209
P-Value [Acc > NIR] : 1.48e-08

Kappa : 0.9274

According to the study methodology, the analysis and discussions of the results, the following were suggested:

- The proposed DEA-DT model performed very well in predicting the efficiencies of bank branches in both cases as it was able to predict in both cases at accuracy of not less than 90% which is on the higher side compared to similar studies like (Wu, 2006).
- The results also suggest that the DEA-DT model works best on predicting bank investment efficiencies as compared to deposit efficiencies.
- The results also suggest that under both deposit and investment stage, the DEA-DT model performs well in both Case 2 and Case 4 with RMSE of 0.299253 and 0.14926

respectively.

4.5.2 RESULTS AND DISCUSSIONS OF THE BANK BRANCHES (132 DMUS) EFFICIENCY ANALYSIS USING THE DEA-RF ALGORITHM PREDICTIVE MODEL

4.5.2 The Bank branches (132 DMUs) efficiency analysis using the RF model Result

In each there were four dependent variables (Classes 1 to 4) and six (6) independent variables for both Case 2 and Case 4. Therefore, the target variables as performance (efficiency) measure

of each category were predicted as either is in Class 1, Class 2, Class 3 or Class 4. To evaluate the prediction of the RF model, k-fold cross validation was implemented to lessen the bias of sampling data and ensuring model error randomness. K-fold cross validation randomly divides data into k subsets and one subset is used as testing data and k-1 subsets are used as training data (Yousaf, 2016) . This process is repeated k times to cover all data. The study used 10-fold Cross Validation in the analysis. In addition to the K-fold cross validation the study also used repeated cross validation where data partitioning was done to split the dataset into training (70%) and testing (30%). To estimate the model's performance, the following Machine Learning performance metrics were used RMSE, MAE, MAPE,

RMSPE and rBIAS. Analysis of the prediction by the DEA-RF model was done in two cases (Case 2 and Case 4). In the first case, Case 2 variables were the dependent variables and the result was compared with the classes designated by the CCR during the DEA stage. The entire results of the RF prediction in both cases are shown in appendix E and F.

4.5.2.1 Random Forest Order of Significant Predictor Variables

During the Case 2 DEA-RF model building and using the permutation variable importance algorithm proposed by Cutler et al. (2011), the DEA-RF estimated the following as important variables with respect to MeanDecraeseAccuracy (shown in Appendix I) in building the model which has been listed in order of magnitude ,thus, most important to the least :

Number of Employees, Total Deposits, IT Expenditure, Percentage of Performing Loans, Fixed Assets and Profits .

For MeanDecraeseGini also shown below in Appendix I, the order is as follows: Number of Employees, Total Deposits, IT Expenditure, Percentage of Performing Loans, Fixed Assets and Profits.

With respect to Case 4 and variable importance according to both MeanDecraeseAccuracy and MeanDecraeseGini are shown in the Appendix I. What was evidence in both cases was the variable Number of Employees and IT Expenditure were always leading the chart in both cases.

4.5.2.2 Case 2-The Deposit Stage

The out-of-bag estimates of the error rate given by $E_{oob} = \frac{1}{N} \sum_{i=1}^N l(y_i, \hat{y}_i)$ were used to choose the optimum Random Forest parameters (formula = Case2 ~ data = TrainSet, ntree = 500, mtry = 2, importance = TRUE). According to the result, the DEA-RF model gave a high overall prediction accuracy of 94.78%. This means that the proposed Random Forest model was able to predict the efficiency of the new 30% bank branches at an overall accuracy rate of about 95% in the Case 2.

The results show that DEA-RF model was able to predict 93 out of 98 (94.9 % accuracy) bank branches that were able to realize overall efficiency even though they were not efficient in collecting deposits from customers. The DEA-RF model was able to predict (92.3%) bank branches in class 3. Finally, the model was able to predict at an accuracy of (95.6%) 22 out of the 23 banks inefficient in collecting deposit and also inefficient in overall banking operations.

The following is also the confusion matrix generated by Case 2 analysis:

```
randomForest(formula = Case2 ~ ., data = TrainSet, ntree =
500, mtry = 6, importance = TRUE)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 6
```

	Reference			
Prediction	Class 2	Class 3	Class 4	Class
2	93	1	1	
Class 3	3	12	0	Class
4	2	0	22	

Overall Statistics

```
Accuracy : 0.9478 95% CI
: (0.8953, 0.9787)
No Information Rate : 0.7313
P-Value [Acc > NIR] : 8.591e-11
```

```
Kappa : 0.8813
```

4.5.2.3 Case 4 –The Investment Stage

For Case 4 out-of-bag estimates of the error rate(E_{oob}) were used to select the optimum Random Forest parameters (formula = Case4 ~., data = TrainSet, ntree = 500,mtry = 6, importance = TRUE). Under the Case 4, the proposed RF model had the highest prediction accuracy of 97.01% as compared to the Case 2 which had 94.78%.

In the case of the Classes 1 and 3, there was no bank branch that was in the test dataset. The DEA-RF model was able to predict 110 out of 113 at an accuracy of 97.3% of banks that realized overall efficiency even when they were not efficient in investing deposit from customers.

Finally, it was able to predict 20 out of 21 bank branches, class 4 –banks efficient in investing deposits and cannot also achieve overall efficiency. The following is also the confusion matrix generated by the Case 4 analysis:

```
randomForest(formula = Case4 ~ ., data = TrainSet, ntree =
500,          mtry = 6, importance = TRUE)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 6
```

Confusion Matrix and Statistics

		Reference			
Prediction	Class	1	Class 2	Class 3	Class 4
1	0	0	0	0	0
Class 2	0	110	0	1	
Class 3	0	0	0	0	
Class 4	0	3	0	20	

Overall Statistics

Accuracy : 0.9701
95% CI : (0.9253, 0.9918)
No Information Rate : 0.8433
P-Value [Acc > NIR] : 2.186e-06
Kappa : 0.8913

According to the study methodology, the analysis and discussions of the results, the following were suggested:

The proposed DEA-RF model performed very well in predicting the bank branches efficiency in both cases as it was able to predict in both cases at accuracy of not less than 94 %.

Finally the results also suggest that the DEA-RF model works best on all the classifications with RMSE of 0.228558 and 0.167248 in Case 2 and Case 4 respectively which is also on a better side compared to similar study by (Hamad & Anouze, 2015).

4.5.3 EMPIRICAL RESULTS AND DISCUSSIONS OF THE PREDICTIONS BY THE DEA-NN MODEL

The study constructed 3-layer networks. The number of hidden neurons (5) based on the equation $N_h = N_i - 1$ proposed by Tamura & Tateishi (1997) were used in both cases and were set to be the same as the back propagation learning algorithm which has been frequently used in business classification studies was adopted. The study also used the “neuralnet” package and implemented it in R (Fritsch et al., 2016) studio using R codes.

To evaluate the prediction of the model, K-fold cross validation was implemented to lessen the bias of sampling data and ensuring model error randomness. K-fold cross validation randomly

divides data into k subsets and one subset is used as testing data and $K-1$ subsets are used as training data (Yousaf, 2016). This process is repeated k times to cover all data. The study used 10-fold cross validation in the analysis. In addition to the k -fold cross validation the study also use repeated Cross Validation where data partitioning was done to split the dataset into training (70%) and testing (30%). To estimate the model's performance, machine learning models such as RMSE, MAE, MAPE, RMSPE and rBIAS were used as the evaluating measures. A 10-fold cross-validation (CV) was applied to check the performance of all predicting models in each case. Analysis of the prediction by the DEA- NN model was done in two cases (Cases 2 and 4). In the first case, Case 2 variables were the dependent variables and the result was compared with the classes designated by the CCR during the DEA stage. The entire results of the DEANN prediction both cases are shown in G and H.

4.5.3.1 The Bank Branches (132 DMUs) Efficiency Analysis Using the DEA-NN Model

Four hundred and forty-four (444) commercial bank branches in Ghana were involved in this Neural Network study where 70% banks branches dataset were used to train or build the DEA-NN model. The proposed NN model was used to predict the efficiency of the remaining 30 % bank branches. The input and output variables used as mentioned earlier were adopted from the dataset used in the previous section during the constructing of the decision tree model and random forest model. Specifically, the cases 2 and 4 were considered.

For the performance analysis of the proposed DEA-NN model, the 30% bank branches dataset normally called the test data were used. Analysis and prediction of banks' efficiencies by the DEA-NN model was done under two separate scenarios (Case 2 and Case 4). In both cases (Case 2 and Case 4), the DEA-NN predicted scores (Classes) were compared with the efficiency classes designated by the CCR during the DEA. It is also important to mention that for the purpose of this study, efficient unit is a unit with an efficiency score of 0.8 or more.

The entire results of the DEA-NN prediction in both cases are shown in Appendices G and H.

4.5.3.2 Case 2 –The Deposit Stage

When the proposed DEA-NN model which was trained using 310 (70%) bank branches dataset was used on the other 134 bank branches, the DEA-NN model was able to predict the efficiencies and their respective classes assigned by the CCR during the DEA at an accuracy rate of 71.64%. The number of inputs used in the case 2 was 6 with 5 hidden neurons on a 3 layer network. The main expected outputs for this scenario were Classes 1 to 4. With respect to Case 2 test dataset, there was no bank branch that has the class 1 characteristics (banks efficient in investing deposit but was not able to attain overall efficiency).

There were 103 banks with the class 2 characteristics (banks that realized overall efficiency even though they were not efficient in investing deposit), 14 banks branches with the class 3 characteristics (Banks that were efficient in investing deposit and still had overall efficiency) and finally 17 banks with the class 4 characteristics (Banks inefficient in investing deposit and cannot also achieve overall efficiency).

According to the findings, DEA-NN model predicted 96 out 103 bank branches realized overall efficiency even though they were not efficient in investing deposit Class 2. This means that the DEA-NN model was able to detect banks in the class 2 at accuracy of 93.2%. For banks in the class 3 and class 4 categories, the DEA-NN model was not able to predict them.

The following is also the confusion matrix of the DEA-NN analysis:

Confusion Matrix and Statistics

	Reference		
Prediction	2	3	4
2	96	14	13
3	7	0	4
4	0	0	0

Overall Statistics

Accuracy : 0.7164179 95%
CI : (0.632137, 0.79087)
No Information Rate : 0.7686567

P-Value [Acc > NIR] : 0.935099793

Kappa : 0.0079875

4.5.3.3 Case 4 –The Investment Stage

Under Case 4, the proposed DEA-NN model had the highest prediction accuracy of 80.59% as compared to Case 2. There were a total of 134 bank branches with the following classes distributions:

- Class 1 had only 1 bank branch
 - Class 2 had the highest number of bank branches of 110
 - Class 3 had only 1 bank branch
 - Class 4 was second highest (22) in terms of the number of bank branches in the dataset
- For both class 1 and class 3, the DEA-NN model predicted all wrongly. This implies that for both classes 1 and 3, the DEA-NN model accuracy was 0% and this may be attributed to the fact that there was not enough Class 1 and Class 3 banks in the training dataset. It was therefore very difficult for the model to predict a variable whose characteristic is not well known and has not been properly learned by machine properly. This is also consistent with the definition of Machine learning given by Alpaydm (2010) - programming computers (Machines) to optimize a performance criterion using example data or past experience (Alpaydm, 2010). In this case, there was no better past experience of both classes 1 and 3 by the proposed DEA-NN model.

With regard to the 110 banks that obtained overall efficiency in their business operations, even though they were not efficient in investing the deposit collected at the stage I of the DEA model, the proposed DEA-NN model was able to detect or predict 108 out of 110 (accuracy rate of

98.18%) of these bank branches. This means that the DEA-NN model performs best when predicting banks that can achieve overall efficiency in their business operations even though they cannot have good efficiency score in investing their deposits or giving loans to customers.

Finally, the DEA-NN model predicted all the other classes of banks wrongly. The following is also the confusion matrix generated by the DEA-NN model:

Confusion Matrix and Statistics

Prediction	Reference			
	1	2	3	4
1	0	0	0	0
2	0	108	1	21
3	1	2	0	1
4	0	0	0	0

Overall Statistics

Accuracy : 0.8059701
 95% CI : (0.7287708, 0.8691644)
 No Information Rate : 0.8208955 P-Value
 [Acc > NIR] : 0.7188772

Kappa : 0.0460022

According to the study, the analysis and discussions of the DEA -NN prediction results, the following were suggested:

- The DEA-NN model performance was very low in predicting the efficiencies of bank branches in both cases as compared to the other two models, thus DEA-DT and DEA-RF.
- Finally the DEA-NN model for classifications suggested MAPE of 24.6% and 20.5% for stage 1 and stage 2 respectively. In the case of the stage 1 classification of the testing data set, the proposed DEA-NN model performed better than the Kwon & Lee (2015) BPNN model with MAPE of 36.9 % for their testing data set.

4.5.4 EMPIRICAL RESULTS AND DISCUSSIONS OF THE PREDICTIONS BY

THE DEA-LR MODEL

To evaluate the prediction of the model, K-fold cross validation was implemented to lessen the bias of sampling data and ensuring model error randomness. K-fold cross validation randomly divides data into k subsets and one subset is used as testing data and K-1 subsets are used as training data (Yousaf, 2016) . To estimate the model's performance, machine learning models such as RMSE, MAE, MAPE, RMSPE and rBIAS were used as the evaluating measures. Analysis of the prediction by the DEA- LR model was done in two cases (Cases 2 and 4). In the first case, Case 2 variables were the dependent variables and the result was compared with the classes designated by the CCR during the DEA stage.

4.5.3.1 The Bank Branches (132 DMUs) Efficiency Analysis Using the DEA-LR Model

Analysis and prediction of banks' efficiencies by the DEA-LR model was done under two separate scenarios (Case 2 and Case 4). In both cases (Case 2 and Case 4), the DEA-LR predicted scores (Classes) were compared with the efficiency classes designated by the CCR during the DEA. It is also important to mention that for the purpose of this study, efficient unit is a unit with an efficiency score of 0.8 or more.

4.5.3.2 Case 2 –The Deposit Stage

When the proposed DEA-NN model which was trained using 310 (70%) bank branches dataset was used on the other 132 bank branches, the DEA-LR model was able to predict the efficiencies and their respective classes assigned by the CCR during the DEA at an accuracy rate of 83.33%. According to the findings, DEA-LR model predicted 110 out 132 bank efficiency. The details of the actual class and predicted class by the DEA-LR model are shown in Appendix L. The following are the Coefficients, Std. Errors, Residual Deviance, PVALUES and ODDS RATIOS realized from the model development used for the predictions.

Call:

```
multinom(formula = Class ~ EMPLOYEES + IT + ASSET, data = data1)
```

Coefficients:

(Intercept) EMPLOYEES IT ASSET

Std. Errors:
 (Intercept) EMPLOYEES IT ASSET
 Class 3 3.044426e-13 2.191791e-12 4.783371e-07 5.600138e-08
 Class 4 1.567412e-13 1.281392e-12 4.611222e-07 1.083685e-07

Residual Deviance:
 304.1911 AIC:
 320.1911

P-VALUES:
 (Intercept) EMPLOYEES IT ASSET
 Class 3 0 0 0.2009169 0.1214307
 Class 4 0 0 0.8159660 0.8630356

ODDS RATIOS:
 (Intercept) EMPLOYEES IT ASSET
 Class 3 3.135161e-03 1.697198 0.9999994 1
 Class 4 1.533586e-13 38.065273 0.9999999 1

4.5.3.3 Case 4 _The Investment Stage

Under Case 4, the proposed DEA-LR model had the highest prediction accuracy of 92.42% as compared to Case 2 which also recorded 83.33%. This means that the DEA -LR model build under the investment stage. Per the results of the analysis, the DEA -LR model built in the investment stage predicted 122 out 132 bank efficiency correct. The details of the prediction realized by the DEA-LR in the investment stage are also shown in Appendix L.

Class 3 -5.765075 0.5289785 -6.117662e-07 8.673473e-08
 Class 4 -29.505998 3.6393024 -1.073200e-07 -1.869479e-08

Compared to the DEA-LR model built in the deposit stage, the DEA-LR model in the investment stage was better as the results of almost all the Machine Learning Evaluation Metrics favored the investment stage model. The following information is also the Coefficients, Std. Errors, Residual Deviance, P-VALUES and ODDS RATIOS of the DEA-LR model in the investment stage.

Call:
`multinom(formula = Class ~ EMPLOYEES + IT + ASSET, data = data1)`

Coefficients:

	(Intercept)	EMPLOYEES	IT	ASSET
Class 2	44.59138	-4.611388	-7.521170e-08	-7.398283e-07
Class 3	44.57503	-5.296252	-2.736245e-06	-8.076129e-07
Class 4	15.21849	-1.044336	-3.581966e-07	-6.803632e-07

Std. Errors:

	(Intercept)	EMPLOYEES	IT	ASSET
Class 2	6.093456e-12	3.853595e-11	2.239300e-06	4.191700e-07
Class 3	7.205787e-12	4.876917e-11	5.527349e-06	5.460845e-07
Class 4	9.330176e-13	9.328377e-12	2.247739e-06	4.190773e-07

Residual Deviance:
129.7539 AIC:
153.7539

P-VALUES:

	(Intercept)	EMPLOYEES	IT	ASSET
Class 2	0	0	0.9732064	0.07756643
Class 3	0	0	0.6205737	0.13916291
Class 4	0	0	0.8733863	0.10448700

ODDS RATIOS:

	(Intercept)	EMPLOYEES	IT	ASSET
Class 2	2.321626e+19	0.009938015	0.9999999	0.9999993
Class 3	2.283956e+19	0.005010337	0.9999973	0.9999992
Class 4	4.067313e+06	0.351925464	0.9999996	0.9999993

CHAPTER 5

GENERAL DISCUSSION

5.0 INTRODUCTION

This study has explored the application of Machine Learning Algorithm in the study of the efficiency of banks using banks in Ghana as a case study. The study proposed a framework called Banks' Efficiency Prediction. This framework was used to build the four different models namely DEA-DT, DEA-RF, DEA-NN and DEA-LR for predicting the efficiency of banks. These combined models were empirically evaluated using real world dataset from a developing country; Ghana and their performances were assessed using RMSE, MAE, MAPE, RMSPE and rBIAS as model performance metrics.

The study started with the development of the four Machine Learning Algorithms models, Decision Tree, Random Forest, Neural Networks and Logistic Regression which were used to classify the banks in Ghana using two scenarios of (Case 2 and Case 4). This was followed by the determination of the various bank efficiencies using a two- stage Data Envelopment Analysis (DEA). Predictor variables were used together with the various efficiency classification classes of the bank branches as the dataset. The DEA scores were classified under four different scenarios using the proposed Banks' Classification Algorithm (BC Algorithm) and were considered as the response variables. Four models were developed and proposed by combining DEA with four Machine Learning Algorithms, Decision Tree, Random Forest, Neural Networks and Logistic Regression. These proposed models namely; DEA-DT, DEA-RF, DEA-NN and DEA-LR were used to predict the efficiency classes of the bank branches using two scenarios of the DEA scores (Case 2 and Case 4). The proposed models' classification results were compared with the classification results of the original Decision

Tree, Random Forest, Artificial Neural Network and Logistic Regression using RMSE, MAE, MAPE, RMSPE and rBIAS as Machine learning performance metrics.

In this last but one chapter, the study generally discusses the results realized from the study, the study's contribution to knowledge and a list of publications available at the time of writing this thesis.

5.1 ANALYSIS OF THE MACHINE LEARNING ALGORITHMS

To estimate the performance of the four proposed models used for the classifications in the study, RMSE, MAE, MAPE, RMSE and rBIAS were considered as the evaluating measures.

A 10-fold Cross-Validation (CV) was also applied to check overfitting and performance of all predicting models, Decision Tree (DT), Random Forest (RF), Neural Network (NN) and Logistic Regression (LR) on each case dataset. The values of the five machine learning evaluation metrics for each of the model are given in the tables below.

For the case 2 dataset, the proposed models namely ; DEA-DT, DEA-RF, DEA-NN and DEALR by the study performed better compared to the Decision Tree, Random Forest, Neural Network and Logistic Regression with DEA-RF performing best with MAPE of 5.22% followed by DEA-DT 8.95% and finally the DEA-LR with 16.67%. The DEA-ANN was the worst performing model in the deposit stage with MAPE of 24.59%.

Table 5.1: Performance of the Models Using Machine Learning Evaluation Metrics under the Deposit Stage

CASE 2- Deposit Stage					
MODELS	RMSE	MAE	MAPE	RMSPE	rBIAS
DT	0.70181	0.492537	49.2537313	0	0.492537
RF	0.64646	0.41791	41.79104	0	0.41791
ANN	1.176902	1.042135	57.45151114	1.179884	-0.019427972
LR	0.583874	0.340909	34.09091	0.583874	1
PROPOSED MODELS					
DEA-DT	0.299253	0.089552	8.955223881	0	0.089552
DEA-RF	0.228558	0.052239	5.223880597	0	0.052239
DEA-ANN	0.685259	0.599699	24.59171999	0.707712	-17.5888
DEA-LR	0.408248	0.166667	16.66667	0.408248	1

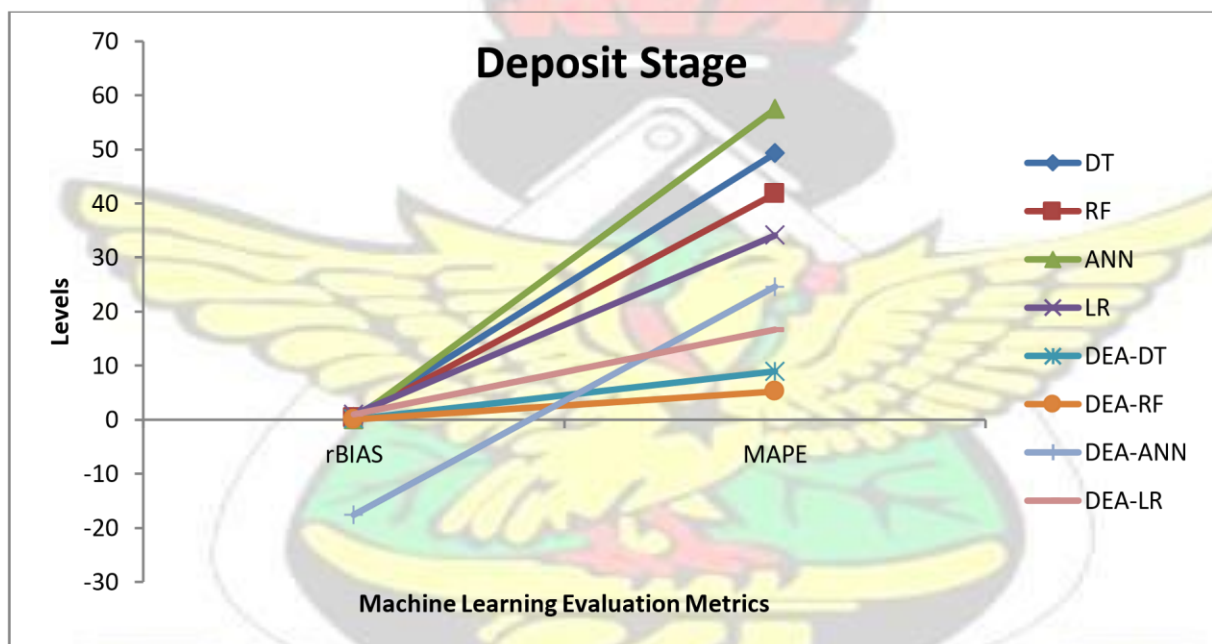
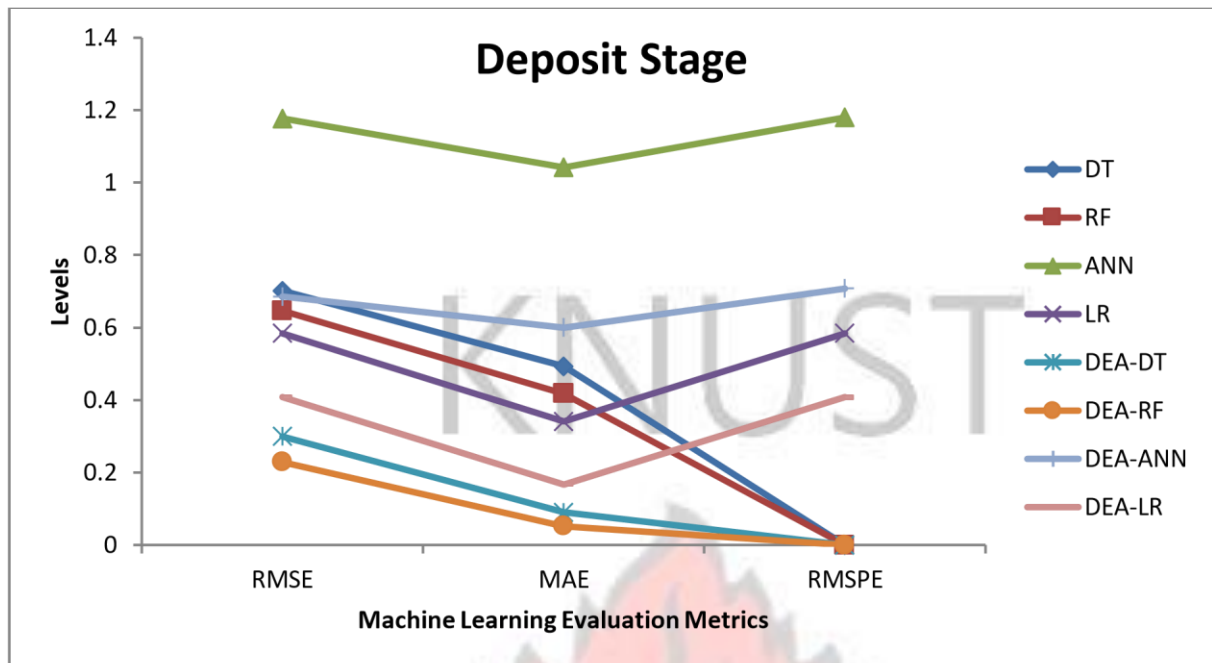


Figure 5.1: Graph showing the Performance of the Machine Learning Algorithms models and the four Models proposed in Case 2 (Author's construct).

Figure 5.1 shows the plot for both Machine Learning Algorithms models and the proposed model under the deposit stage. The results did not show any level of good performance of the machine learning algorithms models compared to the proposed models by the study which is an indication of a better machine learning model. Without exceptions, all the models estimates are significant at the 0.05 level of significance. The study results also reveal that among the

four proposed models (DEA-DT, DEA-RF, DEA-NN and DEA-LR) for predicting the efficiency of banks at the deposit stage; the DEA-RF is suitable for modeling bank deposit efficiency data. As it was the model with the best RMSE, MAE, MAPE, RMPSE and rBIAS. Also, by inspecting the MAE values which is a measure of prediction accuracy, we observe that MAE values for proposed models are smaller than those for the DT, RF, NN and LR models. Hence the proposed model is selected for deposit efficiency analysis and classification of banks. The proposed framework is better when comparing its performance with a similar study by Hamad & Anouze (2015) where the authors introduces a three-stage integrated framework consisting of data envelopment analysis (DEA), random forest, and logistic regression to examine and predict the impact of environmental variables on banks' performance and had 79.4% accuracy rate of classification using 151 banks in Middle East and North African (MENA) countries.

Table 5.2: Performance of the Models Using Machine Learning Evaluation Metrics under the Investment Stage

CASE 4-Investment Stage					
MODELS	RMSE	MAE	MAPE	RMSPE	rBIAS
DT	0.782266	0.61194	61.19403	0	0.61194
RF	0.772667	0.597015	59.7015	0	0.597015
ANN	1.241123	1.11657	65.49859	1.227733	-0.02396
LR	0.758787	0.575758	57.57576	0.758787	1
PROPOSED MODELS					
DEA-DT	0.149626	0.022388	2.238806	0	0.022388
DEA-RF	0.167248	0.02985075	2.98507463	0	0.029851
DEA-ANN	0.647185	0.474046108	20.53436629	0.677297	-0.05259
DEA-LR	0.275241	0.075758	7.575758	0.275241	1

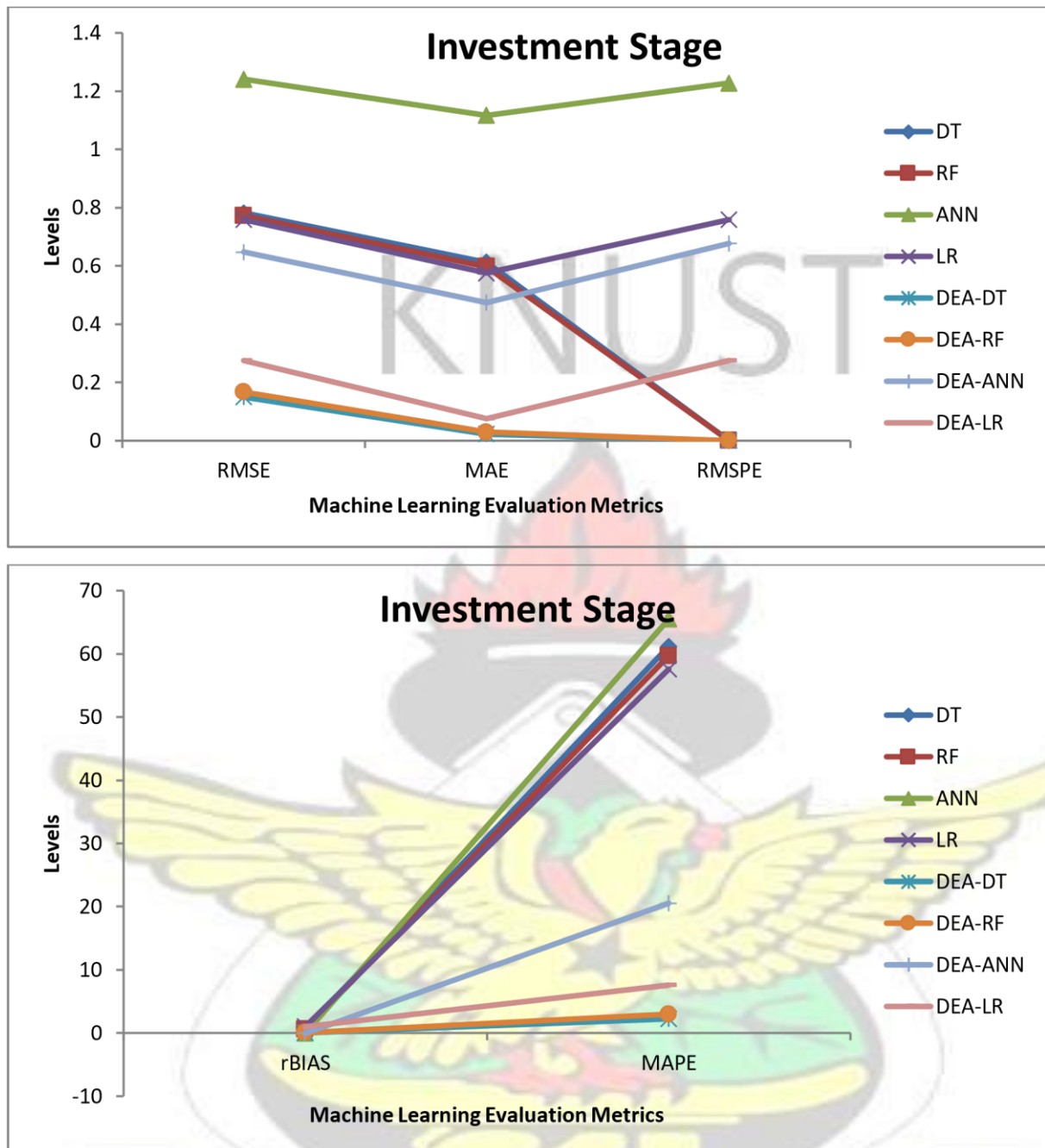


Figure 5.2: Graph showing the Performance of the Machine Learning Algorithms models and the four Models proposed in -Case 4 (Author's construct)

Figure 5.2 shows the plot for both Machine Learning Algorithms models and the proposed model under the investment stage. The results did not show any level of good performance of the machine learning algorithms models (Decision Tree, Random Forest, Neural Network and Logistic Regression) compared the proposed models by the study which is an indication of a better machine learning model. Without exceptions, all the models estimates are significant at

the 0.05 level of significance. The study results also reveal that among the four proposed models (DEA-DT, DEA-RF, DEA-NN and DEA-LR) predicting the efficiency of banks at the deposit stage; the DEA-DT is suitable for modelling bank investment efficiency data. This is because; it was the model with the best RMSE, MAE, MAPE, RMPSE and rBIAS. Also, by inspecting the MAE values which is a measure of prediction accuracy, we observe that MAE values for proposed models are smaller than those for the DT, RF NN and LR models. Hence the proposed model is selected for deposit efficiency analysis and classification of banks. The proposed framework is better when comparing its performance with a similar study by Wu (2006) where the author detected the impact of IT on firm performance by proposing a generic model using Data Envelopment Analysis (DEA) Decision Trees (DTs) and 36 Canadian banks as DMUs.

5.3 Discussion

The assessment of banks efficiencies creates serious problems for banks and their managers (Adusei, 2016; Sufian et al., 2016 and Titko, Stankevičienė, & Lāce, 2014). Due to the critical significant role of banks in the national economy, their efficiencies and performances are a hotly debated topic in the academic and business environments (Titko et al., 2014). Models and frameworks designed for this assessment are normally based on econometric analysis (Brynjolfsson, 1993; Dedrick et al., 2013; Dedrick, Gurbaxani, & Kraemer, 2003; Kılıçaslan, Sickles, Kayış, & Gürel, 2017; Nations, 2008 and Roghieh, Saeed, & Sang-Yong, 2004) and other parametric methods.

Non-parametric techniques such as Data Envelopment Analysis (DEA) have been suggested in literature as technique to evaluate bank efficiency (Lampe & Hilgers, 2015). However, such studies (Chen et al., 2006; Chen & Zhu, 2004a; Izadikhah, Tavana, & Di, 2017; Madjid et al., 2009; Tavana, Izadikhah, Caprio, & Saen, 2017; Wang, Gopal, & Zionts, 1997) that used only non-parametric methods such as pure DEA and its models suffer from weak discrimination power (Ashoor, 2012; Chen, 2016; Santos et al., 2017; Wu, 2006, 2009) and also sensitive to the presence of outliers and statistical noise (Da, Stasinakis, & Bardarova, 2018 and Dash et

al., 2006). It is also very difficult to use only DEA to predict the efficiency and performance of other or new Decision Making (Emrouznejad & Yang, 2017; LaPlante & Paradi, 2014 and Wanke et al., 2016).

Combined DEA and Machine Learning Algorithms can offer these suitable and scientific methods. A lot of machine learning algorithms have been developed and utilized in the business and financial sectors for predictions and forecasting. For instance, Chen & Hao (2017) used Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for stock market indices prediction using the two well-known Chinese stock market indices. There are other studies such as Abellán & Castellano (2017); Ahn, Cho, & Kim (2000); Anandarajan, Lee, & Anandarajan (2001); Caggiano, Calice, & Leonida (2014); Caggiano, Calice, Leonida, & Kapetanios (2016); Göçken, Özçalıcı, Boru, & Ayse (2016); Janek, Beyers, Pieter, & Villiers (2016); Kim & Han (2000); Patel, Shah, Thakkar, & Kotecha (2014b, 2014a); Qiu, Song, & Akagi, (2016) and Tsai & Wu, (2008) that have implemented machine learning algorithms in the business and financial sectors in forecasting and prediction studies.

According to literature, Machine Learning Algorithms used to build predictive models have been suggested as the one of finest and best predicting methods with a high accuracy and validity in the field business and financial forecasting (Göçken et al., 2016). Nevertheless, prediction and classification of bank branches using with a high performance models across developing countries are also completely lacking in literature (Mohd Zaini Abd, 2001 and Tra et al., 2018) According to Yang et al. (2015) Zheng et al. (2004) and validated by the findings in this study combined models consistently outperform individual models for classification and prediction tasks. This current work uses a primary data gathered from more than four hundred (400) DMUs of bank branches.

In comparison with other studies such as Hamad & Anouze (2015); Kwon & Lee (2015) and Wu (2006), the novelty of this thesis lies in both the significance of the frameworks /models

(DEA-DT, DEA-RF, DEA-NN and DEA-LR) applied in the study ,the algorithm proposed and used and the key findings observed. First and foremost, this study combines a two-stage DEA model with different Machine Learning Algorithms namely Decision Tree, Random Forest, Artificial Neural Network and Logistic Regression, where the efficiency at each stage (thus efficiency in collecting deposit from customers, Deposit efficiency and efficiency in investing the deposit, Investment efficiency) is also considered as input for the next stage using a huge dataset a from a developing country. The study also classified the various banks based on their efficiency scores using a proposed algorithm, Bank Classification Algorithm suggested by the study. This efficiency classes (Class1, Class2 Class 3 and Class 4) were used as the response variable making, the response variable for the study categorical. For the development of the four models, the study considered predictor variables which were both internal and external and directly influence bank performance and efficiency as suggested by (Thanassoulis et al., 2008). The development and performance of the proposed in this current study was done on big dataset compared to Wu (2006) work which was done on a small dataset from existing literature (Wang et al., 1997). Kwon & Lee (2015) also used a total of 181 DMUs as a dataset. Which accooidng to this same Kwon & Lee (2015), a large data set is often favored for the generalization of models. A K-fold cross-validation was used to better the performnace of the models as compared to Wu (2006) works, which was based on V-fold cross validation. The data used in this study is not a cross-country bank-level data compared to Hamad & Anouze (2015). In this study ,70% of the data set (444 DMUs) was used for training and 10 % validation and 30% for testing the models compared to Kwon & Lee, (2015) study of combining DEA with Back Propagation Neural Network which was based on a ratio of 60:20:20 for training,validating and testing respectively. This is also a single study that has build and proposed four different models that has yieded fovorable classification accuracy rate. These models were developed by combining DEA with Decision Tree (DEADT),Random Forest (DEA-RF), Artificial Neural Network (DEA-NN) and Logistic Regression (DEA-LR). In the case of the DEA-DT, the model perfomed better than Wu

(2006) work by giving an accuracy of not less than 90% compared to Wu (2006) 69.44%. Comparing the DEA-RF to that of Hamad & Anouze (2015), the study also suggested a favourable classification accuracy of 90% compared to Hamad & Anouze (2015) work that also yielded 79.4% and finally, the DEA-NN suggested a favourable Mean Absolute Percentage Error (MAPE) of about 24.6% compared to the BPNN MAPE of 36.9% also suggested by (Kwon & Lee, 2015). Compared to the works of Hamad & Anouze (2015); Kwon & Lee, (2015) and Wu (2006), this study demonstrated the performance of the four proposed models by using five standard machine learning evaluation metrics namely; Root Mean Square Error (RMSE), Mean Absolute Performance Error (MAPE), Mean Absolute Error (MAE), Root Mean Square Performance Error, and rBIAS.

Even though it has been noted by earlier authors that combining DEA with machine learning algorithms will yield favorable results (Wu, 2006) in terms of classification. An important observation that can be made from the studies conducted is that most of these frameworks were not able to produce machine learning algorithm models that can suggest a higher classification accuracy of more than 80%. Comparing the results of four machine learning models which were built by combining DEA with machine learning algorithms in a single study using a bigger data set (444 DMUs) is very scarce in literature. This thesis therefore contributes significantly to filling this particularly huge gap in the literature of bank efficiency prediction.

5.4 ORIGINALITY AND CONTRIBUTIONS TO KNOWLEDGE

It is well acknowledged that making an original contribution to knowledge is often a contentious issue among scholars, especially in doctoral studies due to the arbitrary nature of the concept of originality. However, there is no doubt that, this study has provided an insight and significant contribution to the body of knowledge as regards assessing efficiency and performance of firms using an improved machine learning model. In comparison with related studies in combining DEA with Machine Learning Algorithms (Hamad & Anouze, 2015; Kwon & Lee, 2015; Wu, 2006), the novelty of this thesis lies in both the significance of the

frameworks /models (**DEA-DT,DEA-RF, DEA-NN and DEA-LR**) applied in the study ,the algorithm proposed and used and the key findings observed. Even though it has been noted by earlier authors that combining DEA with machine learning algorithms will yield favorable results (Wu, 2006) in terms of classification. An important observation that can be made from the studies conducted is that most of these frameworks were not able to produce machine learning algorithm models that can suggest a higher classification accuracy of more than 80%. Comparing the results of four machine learning models which were built by combining DEA with machine learning algorithms in a single study using a bigger data set (444 DMUs) is very scarce in literature. This thesis therefore contributes significantly to filling this particularly huge gap in the literature of bank efficiency prediction.

In relation to bank efficiency assessment and analysis, the following novel contributions to knowledge were achieved;

1. The development of a high accuracy machine learning models (**DEA-DT, DEA-RF DEANN and DEA-LR**) for predicting bank efficiency by combining DEA and four machine learning algorithms namely; Decision Tree algorithm, Random Forrest algorithm, Artificial Neural Network and Logistic Regression.
2. The development of framework called **Banks' Efficiency Prediction** which has also proved to be effective when it was empirically tested with real world numerical data.
3. The design of a **Bank Classification Algorithm (BC Algorithm)** that was used to classify banks in into classes based on their efficiencies.
4. The used of percentage of non-performing loans as an output variable in a two stage DEA for assessing the efficiency of banks.
5. The use of efficiency value in each of the stages as input variable in the next stage in DEA. Thus the efficiency score for the stage I was used together with the deposits collected as inputs for stage II and finally, the efficiency for stage II was also used with the fixed assets, IT budget and number of employees as input for the overall stage.

6. This study has also contributed to literature with its huge dataset on DMUs.

This study has also contributed to the existing literature through the following journal publication and conference presentation:

Journal Articles

1. Appiahene, P., Missah, Y. M., & Najim, U. (2020). Predicting Bank Operational Efficiency Using Machine Learning Algorithm : Comparative Study of Decision Tree , Random Forest , and Neural Networks. *Advances in Fuzzy Systems*, 2020.
<https://doi.org/https://doi.org/10.1155/2020/8581202>.
2. Appiahene, P., & Missah, Y. M. (2019a). Predicting the Operational Efficiency of Banks in the presence of Information Technology Investment. *International Journal of Advances in Electronics and Computer Science*, 6(11), 1–6. <https://doi.org/IJAECs-IRAJ-DOI-16504>.
3. Appiahene, P., Missah, Y. M., & Najim, U. (2019). Evaluation of information technology impact on bank ' s performance : The Ghanaian experience. *International Journal of Engineering Business Management*, 11, 1–10. <https://doi.org/10.1177/1847979019835337>.
4. Appiahene, P., Ussiph, N., & Missah, Y. M. (2018). Information Technology Impact on Productivity : A Systematic Review and Meta- Analysis of the Literature. *International Journal of Information Communication Technologies and Human Development*, 10(3), 39–61. <https://doi.org/10.4018/IJICTHD.2018070104>.
5. Predicting the Operational Efficiency of Banks using their Information Technology: Decision Tree Algorithm Approach. **Under Review**

Conference Presentations and Proceedings

1. Predicting the Operational Efficiency of Banks in the Presence of Information Technology Investment Using Artificial Neural Network. Presented at the International

Conference on Artificial Intelligence and Soft Computing (ICAISC) Munich, Germany 3rd - 4th July 2019.

2. Appiahene, P., & Missah, Y. M. (2020). Bank Classification Algorithm: Case Study of Ghanaian Banks. 2019 International Conference on Communications, Signal Processing and Networks (ICCSPN), 1–6. <https://doi.org/10.1109/iccspn46366.2019.9150171>.
3. Information Technology Impact on Organizational Productivity: A Survey of the Literature was presented at the 5th International Conference on Applied Sciences and Technology (ICAST) held on September 2018 in Kumasi, Ghana.

CHAPTER 6 6.0 CONCLUSION AND RECOMMENDATIONS

Firstly, against the background of few studies on the assessment of bank efficiencies using data from Ghana, this study has been undertaken in an effort to close the gap in literature between Ghana and the rest of the world. Secondly, there are few studies that have combined DEA with Machine Learning Algorithm like Decision Tree, Random Forest, Neural Network and Logistic Regression to predict the efficiency of banks. These banks were classified based on their efficiency scores at the deposit and investment stages using a proposed Bank Classification Algorithm (BC Algorithm). The study also provided an empirical assessment of four Machine Learning Algorithms using Ghanaian banks as a case study. The evaluation undertaken in this study indicates that combined models consistently outperform individual models for classification and prediction tasks suggested by Yang et al. (2015) and Zheng et al. (2004).

In this study, a framework was proposed; DEA was combined with four different Machine Learning Algorithms to predict banks' efficiencies using Ghanaian banks as a case study. The dependent variables were Case 2 (Deposit efficiency) and Case 4 (Investment efficiency)

efficiency scores classes that were realized from determining the efficiencies of the various bank branches.

Firstly, using existing literatures on the topic, the study identified the most commonly used and cited methodologies used to assess the efficiency and performance of an organization. This resulted in the integration of a non-parametric model - a two-stage DEA which has been proved to be a good measurer of organizational performance or efficiency Chen et al. (2006); Chen & Zhu (2004b) with Machine Learning Algorithm. Data collected from the selected banks, the DMUs were divided randomly into two: training and validation dataset (70%) and testing dataset (30%). The predictor variables were the various input and output parameters that were used to calculate the efficiency scores of each DMU using the DEA. These efficiency scores were classified under the deposit and investment stages. The combination of the DEA scores, classes (response or dependent variable) under each case and predictor variables form the dataset for the study. For the prediction models, the study utilized four popular machine learning algorithms and compared the four algorithms using several performance metrics of algorithm measurements. The best performed Machine Learning Algorithm model in each case, cases 2 and 4 using several performance measures were determined based on the 30% testing dataset. The results of the predictive models indicated the following:

- Combing a non-parametric model like DEA with a machine learning algorithm work better than standalone Machine Learning Algorithm.
- A combination of Data Envelopment Analysis, and Machine Learning Algorithms such as Decision Tree, Random Forest, Neural Network, and Logistic Regression gives high performance accuracy in predictions which confirms Yang et al. (2015) and Zheng et al. (2004) suggestions.
- The best Machine Learning model for predicting the efficiency of banks in terms of collecting deposit from customers is the proposed **DEA-RF**.
- The best Machine Learning model for predicting the efficiency of banks in terms of investing customers deposit is the proposed **DEA-DT**.

6.1 LIMITATIONS OF THE FINDINGS

Ascertaining the precincts of any academic work helps improve its reliability and the generalizability applications of the findings. There are some possible limitations that should be borne in mind in the analysis and generalization of the study findings.

The focus of the numerical illustration aspects of this study was entirely based on the Ghanaian banking industry experience. Given that in practical terms various economic indicators may differ across countries, geographical regions or even continents; it is completely possible that there may be substantial differences and disparities in the findings if this study is simulated in other countries or geographical regions. However, hypothetically, it can be said that the banking industry in many developing countries especially in sub-Saharan Africa are considered to exhibit comparable practical and economic characteristics. Hence, this limitation stated above does not emasculate the validity of the study undertaken and potential application of its main findings in sister developing countries especially those in Africa. Furthermore, the convergence of the findings with general body of knowledge supported by the validation results further reinforces the credibility of the research findings.

It is very important to acknowledge the limitation of the relatively few population sample considered in the study imposes. As the study only considered 444 commercial bank (universal banks) branches in Ghana, expanding the sampling data to include other financial firms like insurance companies, savings and loans, rural banks, credit unions could have enriched the findings and increased its potential generalization. Nevertheless, this should not invalidate the findings and conclusions as the demographic profile of the 444 selected bank branches show a cross fertilization of almost all the other financial institutions in Ghana.

As suggested, most statistical analytical approaches and tools are affected by issues of multicollinearity, sampling inconsistencies, measurement errors, analytical bias which are likely to impact on the results and the potential conclusions to be drawn from the findings. However, notwithstanding, the potential of these highlighted above, it can be suggested that the

demographic profile of the respondents in terms of experience, knowledge & understanding on the topic and consistencies registered in the statistical analysis indicate some degree of reasonable credibility and trustworthiness in the results from the data given.

Finally, with respect to factors that affect bank's deposit collection, there were other several factors that were not taken into consideration because of inadequate data on these factors. Some of these are listed below;

- Environment: Consider 2 banks, one located in Accra central and the other in a small town Sogakope may have the same inputs, but the deposit mobilization will be significantly different.
- Customers: The type of customers you have makes a big difference; small traders as against large businesses etc.
- Sector of operation. Giving loans to farmers will result in more defaulting loans. Loans to contractors not paid on time by the Government have caused defaulting loans even collapsing some banks.
- Bank Policies: This will restrict the amount of loans that can be given and the sectors.
- Large loans: Loans beyond the capacity of the local branch will require assistance and approval from the head office. All these factors can affect deposit mobilization as suggested by (Ambe, 2017; Gunasekara & Kumari, 2018; Jembere, 2016; Joyce, 2013; Madebo, 2013; Mushtaq & Siddiqui, 2017; Nahidul et al., 2019; Ostadi & Sarlak, 2014; Pesa & Muturi, 2015; Turhani & Hoda, 2016; Vuong et al., 2020) in their studies. This study is therefore limited to Bank's Fixed Asset, IT expenditure and Total Number of Employees as factors that can impact deposit mobilization by Banks.

6.2 RECOMMENDATIONS

1. As the empirical testing of the proposed framework was done using dataset from banks in Ghana, these developed models, especially the DEA-DT and DEA-RF is therefore recommend to bank managers and other researchers for predicting and classifying banks.
2. The study recommends the usage of the proposed algorithm to classify banks in Ghana based on the efficiency in doing business like deposit collections and investing the deposits.
3. The Management board of banks and other stakeholders may detect new ways to better their efficiency through the use of innovative technologies using the result of this thesis.
4. The general Ghanaian banking industry and other firms may use the results of this thesis to help them have a better understanding of relationship between their resources and the performance and efficiency of their banks.
5. Customers or investors can use the findings of this study as it would add value to their knowledge of how efficient their various bank branches are in terms of managing their investments.
6. The proposed framework can be used by managers of various firms and researchers to understand and correlate the impact of their resources on their firm performance.
7. To improve banks efficiency and performance in order to remain competitive in this current banking crisis in Ghana managers and stakeholders should not over rely on their overall efficiency or performance as means of measuring their growth, but should critically find ways of improving their efficiency in deposit collections from customers and also investing the deposits. This study suggests that a lot of Ghanaian banks are inefficient in investing customers' deposits.
8. Finally the study recommends to the central bank (Bank of Ghana) to use the method and algorithm proposed in this study to classify the various banks in Ghana based on their efficiency in collecting deposits and investing deposits since these two (deposit and

investing deposits) have become hot and critical issue in the current banking crisis in Ghana (Abubakar, 2018; Addison, 2017, 2018a and Addison, 2018).

6.3 FUTURE STUDIES

- Future studies can also test the proposed framework and the bank classification algorithm using a different dataset from other dual role operating firms such as insurance companies and rural banks.
- The dataset used in this study can be analyzed using a different methodology in terms of how the DEA scores were determined and the results compared with this study.
- Future studies can also consider the liquidity ratio on banks as output variable in the determination of the efficiency of banks.
- Future studies can also incorporate other factors such as Environment, Customers, Sector of operation, Bank Policies and Large loans as inputs for Deposit Stage and compare the results with this current study. Thus how these factors can also impact the model.

LIST OF REFERENCES

- Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Application*, 73, 1–10. <https://doi.org/10.1016/j.eswa.2016.12.020>
- Abri, A. G., & Mahmoudzadeh, M. (2014). Impact of information technology on productivity and efficiency in Iranian manufacturing industries. *Journal of Industrial Engineering International*, 11(1), 143–157. <https://doi.org/10.1007/s40092-014-0095-1>
- Abubakar, I. (2018). *Shareholders, directors of defunct UT, Capital Bank engaged in “willful deceit” - BoG*. Myjoyonline.com. <https://www.myjoyonline.com/business/2018/August-7th/shareholders-directors-of-defunct-ut-capital-bank-engaged-in-willful-deceit-bog.php>
- Addison, Y. E. (2017). *Minimum capital requirement for banks pegged at ₵400 million*. MyJoyOnline.com. <https://www.myjoyonline.com/business/2017/September-8th/minimum-capital-requirement-for-banks-to-reach-400-million.php>
- Addison, Y. E. (2018a). *19 Banks meet new capital requirement*. Graphic Online. <https://www.graphic.com.gh/business/business-news/19-banks-meet-newcapitalrequirement.html>

- Addison, Y. E. (2018b). *Banking crisis: BoG's roadmap for clearing UT, Capital bank mess*. Citinewsroom.com. <https://citinewsroom.com/2018/08/14/banking-crisis-bogsroadmapfor-clearing-ut-capital-bank-mess/>
- Addison, Y. E. (2018c). *BoG collapses 5 banks into Consolidated Bank Ghana Ltd | General News 2018-08-01*. Ghanaweb.com. <https://www.ghanaweb.com/GhanaHomePage/NewsArchive/BoG-collapses-5banksinto-Consolidated-Bank-Ghana-Ltd-673691>
- Addison, Y. E. (2018). *BoG enforces banks' Capital Requirement Directive*. Citinewsroom.com. <https://citinewsroom.com/2018/07/02/bog-enforces-bankscapitalrequirement-directive/>
- Adusei, M. (2016). Modelling the efficiency of universal banks in Ghana. *Quantitative Finance Letters*, 9502(July), 60–70. <https://doi.org/10.1080/21649502.2016.1262938>
- Aggelopoulos, E., & Georgopoulos, A. (2017). Bank branch efficiency under environmental change: A bootstrap DEA on monthly profit and loss accounting statements of Greek retail branches. In *European Journal of Operational Research* (Vol. 261, Issue 3). Elsevier B.V. <https://doi.org/10.1016/j.ejor.2017.03.009>
- Ahn, B. S., Cho, S. S., & Kim, C. Y. (2000). The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert Systems with Application*, 18, 65–74.
- Akena, D., Joska, J., & Stein, D. J. (2018). Sensitivity and specificity of the Akena Visual Depression Inventory (AViDI-18) in Kampala (Uganda) and Cape Town (South Africa). *The British Journal of Psychiatry*, 1–7. <https://doi.org/10.1192/bjp.2018.9>
- Alexander, W. R. J., Haug, A. A., Jaforullah, M., & Haug, A. (2007). *A two-stage doublebootstrap data envelopment analysis of efficiency differences of New Zealand secondary schools*. 714.
- Alinezhad, A. (2016). An Integrated DEA and Data Mining Approach for Performance Assessment. *Iranian Journal of Optimization*, 8(2), 59–69.
- Alpaydm, E. (2010). *Introduction to Machine Learning* (T. Dietterich, C. Bishop, D. Heckerman, M. Jordan, & M. Kearns (eds.); Second Edi). The MIT Press Cambridge, Massachusetts London, England.
- Álvarez, I. C., Barbero, J., & Zofío, J. L. (2016). *Economic Anaysis Working Paper Series :A Data Envelopment Analysis Toolbox for MATLAB* (No. 3; 3/2016).
- Álvarez, R. (2016). *The Impact of R & D and ICT Investment on Innovation and Productivity in Chilean Firms Innovation and Productivity in Chilean Firms*.
- Ambe, M. E. (2017). An Investigation of Determinants of Deposit Mobilization in Commercial Banks of Ethiopia. *Research on Humanities and Social Sciences*, 7(19).
- Ananda, D. B., & Wibisono, A. (2014). C4.5 Decision Tree Implementation in Sistem Informsi Zakat (SIZAKAT) to Automatically Determining the Amount of Zakat Received by Mustahik. *Journal of Information Systems*, Volume 10, Issue 1, April 2014, 10(1), 29–36.

- Anand Prakash, Sanjay Kumar Jha, Kapil Deo Prasad, A. K. S. (2017). Productivity , quality and business performance : an empirical study. *International Journal of Productivity and Performance Management*, 66(1), 78–91. <https://doi.org/10.1108/IJPPM-03-2015-0041>
- Anandarajan, M., Lee, P., & Anandarajan, A. (2001). Bankruptcy Prediction of Financially Stressed Firms : An Examination of the Predictive Accuracy of Artificial Neural Networks. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 81(March 2000), 69–81. <https://doi.org/10.1002/isaf.199>
- Anantwar, S. G., & Shelke, R. R. (2012). Simplified Approach of ANN : Strengths and Weakness. *International Journal of Engineering and Innovative Technology (IJEIT)*, 1(4), 73–77.
- Angel, L., Viola, J., Vega, M., & Restrepo, R. (2016). Sterilization process stages estimation for an autoclave using logistic regression models. *2016 21st Symposium on Signal Processing, Images and Artificial Vision, STSIVA 2016, August*, 1–6. <https://doi.org/10.1109/STSIVA.2016.7743337>
- Anouze, A. L. M. (2019). *Data envelopment analysis and data mining to efficiency estimation and evaluation*. 12(2), 169–190. <https://doi.org/10.1108/IMEFM-1120170302>
- Anthanasopoulos, A. D., & Curram, S. P. (1996). A comparison of Data Envelopment Analysis and Artificial Neural Network as Tools for Assessing the Efficiency of Decision making Units. *Journal of Operational Research Society*, 47, 1000–1016.
- Antonija, B., Hassan, A., & James, K. (2017). Analysis of the Banking Sector Performance in Bosnia and Herzegovina, Montenegro and Serbia before and after the Global Financial Crisis. *Economics*, 5(2), 83–101. <https://doi.org/10.1515/eoik-2017-0029>
- Appiahene, P., & Missah, Y. M. (2019). Predicting the Operational Efficiency of Banks in the presence of Information Technology Investment. *International Journal of Advances in Electronics and Computer Science*, 6(11), 1–6. <https://doi.org/10.1109/IJAECs-IRAJOI16504>
- Appiahene, P., & Missah, Y. M. (2020). Bank Classification Algorithm: Case Study of Ghanaian Banks. *2019 International Conference on Communications, Signal Processing and Networks (ICCSPN)*, 1–6. <https://doi.org/10.1109/iccspn46366.2019.9150171>
- Appiahene, P., Missah, Y. M., & Najim, U. (2019). Evaluation of information technology impact on bank ' s performance : The Ghanaian experience. *International Journal of Engineering Business Management*, 11, 1–10. <https://doi.org/10.1177/1847979019835337>
- Ascarya Diana, Y. (2007). Comparing the Efficiency of Islamic Banks in Malaysia and Indonesia. *Pada International Conference on Islamic Banking & Finance*. <https://doi.org/10.1177/001452468209300802>
- Ascarya Yumanita, D., Achsani, N. A., & Rokhimah, G. S. (2008). Measuring The Efficiency of Islamic Bank in Indonesia and Malaysia Using Parametric and Nonparametric Approach. *3rd International Conference on Islamic Banking and Finance, July*.
- Ashoor, L. A. (2012). *Performance Analysis integrating Data Envelopment Analysis and Multiple Objective Linear Programming* . University of Manchester.

- Avkiran, N. (2006). Productivity Analysis in the Service Sector with Data Envelopment Analysis. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2627576>
- Avkiran, N. K. (2014). An illustration of dynamic network DEA in commercial banking including robustness tests. *Omega*. <https://doi.org/10.1016/j.omega.2014.07.002>
- Azadeh, A., Saberi, M., Tavakkoli, R., & Javanmardi, L. (2011). An integrated Data Envelopment Analysis – Artificial Neural Network – Rough Set Algorithm for assessment of personnel efficiency. *Expert Systems With Applications*, 38(3), 1364–1373. <https://doi.org/10.1016/j.eswa.2010.07.033>
- Bajaj, P., Pandey, M., Tripathi, V., & Sanserwal, V. (2019). Efficient Motion Encoding Technique for Activity Analysis at ATM Premises. *Progress in Advanced Computing and Intelligent Engineering*, 393–402. https://doi.org/10.1007/978-981-13-1708-8_36
- Bani-Hani, J. S., Al-Ahmad, N. M. M., & Alnajjar, F. J. (2009). the Impact of Management Information Systems on Organizations Performance: Field Study At Jordanian Universities. *Review of Business Research*, 9(17), 127–137. <http://ezproxy.library.uvic.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=45462891&site=ehost-live&scope=site>
- Banker, R. D., Charnes, A., & Cooper, W. . (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science*, 30(9).
- Biau, G., Biau, O., & Rouvière, L. (2006). Nonparametric Forecasting of the Manufacturing Output Growth with Firm-level Survey Data The data. *DEEEWorkshop, INSEE, Paris*.
- Bluman, A. G. (2009). *Elementary Statistic : a Step By Step Approach* (9th ed.). McGrawHill Education.
- Booth, A., Gerding, E., & McGroarty, F. (2014). Automated trading with performance weighted random forests and seasonality. *Expert Systems With Applications*, 41(8), 3651–3661. <https://doi.org/10.1016/j.eswa.2013.12.009>
- Boritz, E., & Kennedy, D. B. (1995). Effectiveness of Neural Network Types for Prediction of Business Failure. *Expert Systems With Applications*, 9(4), 503–512.
- Braaten, Ø. (2010). *Artificial intelligence applied to medical genetics*. University of Oslo.
- Bradley, A. P. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition Society*, 30(7), 1145–1159.
- Bradtke, S., & Duff, M. (1995). Reinforcement learning methods for continuous-time Markov decision problems. *Advances in Neural Information Processing*. <http://papers.nips.cc/paper/889-reinforcement-learning-methods-for-continoustimemarkov-decision-problems.pdf>
- Breiman, L. (2001a). *Bagging Trees Random Forests Regression Tree / Classification Tree*. 24(2), 1–14. <https://doi.org/10.1023/A>
- Breiman, L. (2001b). Random Forests. In *Machine Learning* (pp. 5–32).
- Brynjoifsson, E. (1996). *Productivity , Business Profitability , and Consumer Surplus : Three Different Measures of Information Technology Value ^ â€™* 2. June.

- Brynjolfsson, E. (1993). *The Productivity Paradox of Information Technology*. 36(12).
- Burki, A. A., & Dashti, I. M. (2003). Cost Efficiency of Commercial Banks in Kuwait and the Need for Regulatory Reforms. In I. Limam (Ed.), *Challenges and Reforms of Economic Regulation in MENA Countries* (Issue January 2003). The American University in Cairo Press.
- Caggiano, G., Calice, P., & Leonida, L. (2014). Early warning systems and systemic banking crises in low income countries : A multinomial logit approach q. *Journal of Banking and Finance*, 47, 258–269. <https://doi.org/10.1016/j.jbankfin.2014.07.002>
- Caggiano, G., Calice, P., Leonida, L., & Kapetanios, G. (2016). Does the Duration of Systemic Banking Crises Matter ? PT Highlights. *Journal of Empirical Finance*. <https://doi.org/10.1016/j.jempfin.2016.01.005>
- Cao, X., & Yang, F. (2011). Measuring the performance of Internet companies using a twostage data envelopment analysis model. *Enterprise Information Systems*, July 2013, 37– 41. <https://doi.org/10.1080/17517575.2010.528039>
- Cardona, M., Kretschmer, T., & Strobel, T. (2013). ICT and productivity: conclusions from the empirical literature. *Information Economics and Policy*. <http://www.sciencedirect.com/science/article/pii/S0167624513000036>
- Carmona, P., Climent, F., & Momparler, A. (2018). Predicting bank failure in the U.S. banking sector: An extreme gradient boosting approach. *International Review of Economics and Finance*. <https://doi.org/10.1016/j.iref.2018.03.008>
- Chang, K. L., Han, I., & Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions. *Decision Support Systems*, 18, 63–72. Chao, W. (2011). *Machine Learning Tutorial*.
- Charnes, A., Cooper, W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2, 429–444.
- Charter, R. A. (1999). Sample Size Requirements for Precise Estimates of Reliability , Generalizability , and Validity Coefficients. *Journal of Clinical and Experimental Neuropsychology*, 21(4), 559–566. <https://doi.org/10.1076/jcen.21.4.559.889>
- Chen, Y., & Hao, Y. (2017). A Feature Weighted Support Vector Machine and K-Nearest Neighbor Algorithm for Stock Market Indices Prediction. *Expert Systems With Applications*. <https://doi.org/10.1016/j.eswa.2017.02.044>
- Chen, Y., Liang, L., Yang, F., & Zhu, J. (2006). Evaluation of Information Technology investment : A data envelopment analysis approach. *Computers & Operations Research*, 33, 1368–1379. <https://doi.org/10.1016/j.cor.2004.09.021>
- Chen, Y., & Zhu, J. (2004a). Measuring Information Technology's Indirect Impact on Firm Performance. *Information Technology and Management*, 5(1–2), 9–22. <https://doi.org/10.1023/B:ITEM.00000008075.43543.97>
- Chen, Y., & Zhu, J. (2004b). Measuring Information Technology's Indirect Impact on Firm Performance. *Information Technology and Management*, 5(1–2), 9–22. <https://doi.org/10.1023/B:ITEM.00000008075.43543.97>

- Chen, Z. (2016). Evaluation and Analysis of the Regional Innovation Efficiencies Based on DEA and Decision Tree. *Management Science and Engineering*, 5(December), 186–192.
- CHI, D.-J., Yeh, C.-C., & Lai, M.-C. (2011). A hybrid approach of dea, rough set theory and random forests for credit rating. *International Journal of Innovative Computing, Information and Control*, 7(8), 4885–4897.
- Cochran, W. G. (1977). *Sampling Techniques* (Third). John Wiley and Sons Inc.
- Coelli, T., & Rao, D. (2005). Total factor productivity growth in agriculture: a Malmquist index analysis of 93 countries, 1980–2000. *Agricultural Economics*.
<http://onlinelibrary.wiley.com/doi/10.1111/j.0169-5150.2004.00018.x/full>
- Commander, S., & Harrison, R. (2011). ICT and productivity in developing countries: new firm-level evidence from Brazil and India. *Of Economics and Statistics*.
http://www.mitpressjournals.org/doi/abs/10.1162/REST_a_00080
- Corrado, C., Haskel, J., Jona-lasinio, C., & Corrado, B. C. (2014). Knowledge spillovers, ICT and productivity growth. In *Knowledge Spillovers, ICT and Productivity Growth* (Issue IZA DP No. 8274).
http://ideas.repec.org/p/imp/wpaper/14624.html%5Cnhttp://papers.ssrn.com/sol3/papers.cfm?abstract_id=2114913
- Creamer, G. (2009). Using Random Forests and Logistic Regression for Performance Prediction of Latin American ADRS and Banks. *Journal of CENTRUM Cathedra*., 2(1), 24–36. <https://doi.org/10.7835/jcc-berj-2009-0020>
- Creamer, G. (2009). Using Random Forests and Logistic Regression for Performance Prediction of Latin American ADRS and Banks by. *Journal of Centrum*, 2(1), 24–36.
- Cutler, A. (2010). *Random Forests for Regression and Classification*.
- Cutler, A., Cutler, D. R., & Stevens, J. R. (2011). Random Forests. In *Machine Learning* (Issue January). <https://doi.org/10.1007/978-1-4419-9326-7>
- Da, F. F. S., Stasinakis, C., & Bardarova, V. (2018). Two-stage DEA- Truncated Regression : Application in banking efficiency and financial development. *Expert Systems With Applications*, 96, 284–301. <https://doi.org/10.1016/j.eswa.2017.12.010>
- Daisuke, M., & Perez, C. (2017). Forecasting Firm Performance with Machine Learning : Evidence from Japanese firm-level data. *RIETI Discussion Paper Series 17-E-068*.
- Dash, D., Yang, Z., & Liang, L. (2006). Using DEA-neural network approach to evaluate branch efficiency of a large Canadian bank. *Expert Systems with Applications*, 31, 108–115. <https://doi.org/10.1016/j.eswa.2005.09.034>
- Dedrick, J., Gurbaxani, V., & Kraemer, K. L. (2003). Information technology and economic performance. *ACM Computing Surveys*, 35(1), 1–28.
<https://doi.org/10.1145/641865.641866>
- Dedrick, J., Kraemer, K. L., Shih, E., Dedrick, J., Kraemer, K. L., & Shih, E. (2013). Information Technology and Productivity in Developed and Developing Countries. *Journal of Management Information Systems*, 1222(August).
<https://doi.org/10.2753/MIS0742-1222300103>
- Delen, D., Kuzey, C., & Uyar, A. (2013).

- Measuring firm performance using financial ratios : A decision tree approach. *Expert Systems With Applications*, 40(10), 3970–3983. <https://doi.org/10.1016/j.eswa.2013.01.012>
- Development, I. (2005). *Economic and Social Implications of ICT. I.*
- Donges, N. (2019). *Tutorials and explanations about applied Machine Learning*. The Logistic Regression Algorithm.
- Dulá, J. . . (2011). An Algorithm for Data Envelopment Analysis. *INFORMS Journal on Computing*, 23(2), 284–296. <https://doi.org/10.1287/ijoc.1100.0400>
- Duygun, M., Prior, D., Sha-ban, M., & Tortosa-Ausina. (2015). Disentangling the European Airlines Efficiency Puzzle: A Network Data Envelopment Analysis Approach. *Omega*. <https://doi.org/10.1016/j.omega.2015.06.004>
- Ekici, B. B., & Aksoy, U. T. (2009). Prediction of building energy consumption by using artificial neural networks. *Advances in Engineering Software*, 40(5), 356–362. <https://doi.org/10.1016/j.advengsoft.2008.05.003>
- El-habil, A. M. (2012). An Application on Multinomial Logistic Regression Model. *Pakistan Journal of Statistics and Operation Research*, 8(2), 271–291. <https://doi.org/10.18187/pjsor.v8i2.234>
- Emmanuel, T. (1999). Data envelopment analysis and its use in banking. *Institute for Operations Research and the Management Sciences*, 29(3), 1–13. <https://doi.org/10.1287/inte.29.3.1>
- Emrouznejad, A., & Anouze, A. L. (2010). Data envelopment analysis with classification and regression tree – a case of banking efficiency. *Expert Systems The Journal Knowledge Engineering*, 27(4), 231–246. <https://doi.org/10.1111/j.1468-0394.2010.00516.x>
- Emrouznejad, A., & Shale, E. (2009). A combined neural network and DEA for measuring efficiency of large scale datasets. *Computers & Industrial Engineering*, 56(1), 249–254. <https://doi.org/10.1016/j.cie.2008.05.012>
- Emrouznejad, A., & Yang, G. (2017). A survey and analysis of the first 40 years of scholarly literature in DEA: 1978–2016. *Socio-Economic Planning Sciences*. <https://doi.org/10.1016/j.seps.2017.01.008>
- Fallahpour, A., Amindoust, A., Antuchevičienė, J., & Yazdani, M. (2017). Nonlinear geneticbased model for supplier selection: a comparative study. *Technological and Economic Development of Economy*, 23(1), 178–195. <https://doi.org/10.3846/20294913.2016.1189461>
- Fritsch, S., Guenther, F., Suling, M., & Sebastian, M. M. (2016). *Package “neuralnet ”* (1.33).
- Fukuyama, H., & Weber, W. L. (2014). Measuring Japanese bank performance : A dynamic Network DEA Approach. *J Prod Anal*. <https://doi.org/10.1007/s11123-014-0403-1> Gal,
- Y. (2016). *Uncertainty in Deep Learning*. September.
- Gemino, A., & Sauer, C. (2010). Using Classification Trees to Predict Performance in Information Technology Projects performance in information technology. *Journal of Decision System*, November 2014. <https://doi.org/10.3166/jds.19.201-223>

- Geoffrey, K. F. T., & Yau, K. . K. W. (2007). Predicting electricity energy consumption : A comparison of regression analysis , decision tree and neural networks. *Energy*, 32, 1761–1768. <https://doi.org/10.1016/j.energy.2006.11.010>
- Ghana, B. of. (2017). *Bank of Ghana Banking Sector Summary: Vol. 2.1*.
- Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. *Pettern Recognition Letters*, 27, 294–300. <https://doi.org/10.1016/j.patrec.2005.08.011>
- Göçken, M., Özçalıcı, M., Boru, A., & Ayse, T. D. (2016). Integrating metaheuristics and Artificial Neural Networks for improved stock price prediction. *Expert Systems*, 44, 320–331. <https://doi.org/10.1016/j.eswa.2015.09.029>
- Greenidge, K., & Grosvenor, T. (2010). Forecasting Non-Performing Loans in Barbados. *Business, Finance & Economics in Emerging Economics*, 5.
- Grmanová, E., & Ivanová, E. (2018). Efficiency of banks in Slovakia: Measuring by DEA models. *Journal of International Studies*, 11(1), 257–272. <https://doi.org/10.14254/2071-8330.2018/11-1/20>
- Grzybowska, U., & Karwański, M. (2014). Families of Classifiers – Application in Data Envelopment Analysis. *Quantitative Methods in Economics*, XV(2), 94–101.
- Gunasekara, H. U., & Kumari, P. (2018). Factors Affecting for Deposit Mobilization in Sri Lanka. *International Review of Management and Marketing*, 8(5), 30–42.
- Halkos, G. E., & Salamouris, D. S. (2004). Efficiency measurement of the Greek commercial banks with the use of financial ratios: A data development analysis approach. *Management Accounting Research*, 15(2), 201–224. <https://doi.org/10.1016/j.mar.2004.02.001>
- Hamad, I., & Anouze, A. L. (2015). Bank efficiency assessment using a hybrid approach of random forests and data envelopment analysis. *Fifth International Conference on Digital Information Processing and Communications (ICDIPC)*. <https://doi.org/10.1109/ICDIPC.2015.7323026>
- Hamad, I. B., & Anouze, A. L. (2015). Bank efficiency assessment using a hybrid approach of random forests and data envelopment analysis. *IEEE*, 182–189.
- Hamid, N., Ramli, N. A., & Hussin, S. A. S. (2017). Efficiency measurement of the banking sector in the presence of non-performing loan. *AIP Conference Proceedings*, 1795(2017). <https://doi.org/10.1063/1.4972145>
- Han, C., Hsieh, C., Lai, F., Li, X., & Han, C. (2011). Information Technology Investment and Manufacturing Worker Productivity. *Journal of Computer Information Systems*, 4417(August).
- Hatefi, M., & Fasanghari, M. (2014). A DEA-Based Approach for Information Technology Risk Assessment through Risk Information Technology Framework. *The International Arab Journal of Information Technology*, 51–58.
- Havidz, S. A. H., & Setiawan, C. (2015). Bank Efficiency and Non-Performing Financing (NPF) in the Indonesian Islamic Banks. *Asian Journal of Economic Modelling*, 3(3), 61–

79. <https://doi.org/10.18488/journal.8/2015.3.3/8.3.61.79>
- Hinton, G. (2018). *How the backpropagation algorithm works Warm up : a fast matrix-based approach to computing the output from a neural.*
- Hornik, K., Buchta, C., Hothorn, T., Karatzoglou, A., Meyer, D., & Zeileis, A. (2019). *Package “RWeka”* (0.4-41 Title). <http://www.cs.waikato.ac.nz/ml/weka/>
- Hssina, B., Merbouha, A., & Bouikhalene, B. (2016). Predicting L earners ' Performa nce in an E-Learning Platform Based on Decision Tree Analysis. *International Arab Conference on Information Technology (ACIT'2016)*, 1–5.
- Hsu, Y., Hsu, M.-F., & Lin, S.-J. (2016). Corporate Risk Estimation by Combining Machine Learning Technique and Risk Measure. *IEEE ICIS 2016*.
- Hu, Zhiguang Yang, J., & Wang, Shuaiwei Yang, Q. (2016). A Hybrid Modified DEA Efficient Evaluation Method in Electric Power Enterprises. *2016 3rd International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS)*, 283–287.
- Izadikhah, M., Tavana, M., & Di, D. (2017). A novel two-stage DEA production model with freely distributed initial inputs and shared intermediate outputs. *Expert Systems With Applications*, 0, 1–18. <https://doi.org/10.1016/j.eswa.2017.11.005>
- Jagoda, K., Lonseth, R., Lonseth, A., Jagoda, K., Lonseth, R., & Lonseth, A. (2013). *A bottom-up approach for productivity measurement and improvement.* <https://doi.org/10.1108/17410401311329625>
- Jain, R. K., Natarajan, R., & Ghosh, A. (2016a). Decision Tree Analysis for Selection of Factors in DEA : An Application to Banks in India. *Global Business Review*, 17(5), 1– 17. <https://doi.org/10.1177/0972150916656682>
- Jain, R. K., Natarajan, R., & Ghosh, A. (2016b). Decision Tree Analysis for Selection of Factors in DEA: An Application to Banks in India. *Global Business Review*, 17(5), 1162–1178. <https://doi.org/10.1177/0972150916656682>
- Janek, J., Beyers, C., Pieter, J., & Villiers, D. (2016). Systemic banking crisis early warning systems using dynamic Bayesian networks. *Expert Systems With Applications*, 62, 225–242. <https://doi.org/10.1016/j.eswa.2016.06.024>
- Jansson, J. (2016). *Decision Tree Classification of Products Using C5.0 and Prediction of Workload Using Time Series Analysis.* KTH Skolan for Elektro- Och Systemteknik.
- Jardin du, P. (2018). Failure pattern-based ensembles applied to bankruptcy forecasting. *Decision Support Systems*, 107, 64–77. <https://doi.org/10.1016/j.dss.2018.01.003>
- Jembere, K. G. (2016). *Ethiopia, Determinants of Commercial Banks Deposit Mobilization Evidence from Private Commercial Banks in Ethiopia* (Issue February). Addis Ababa University.
- Jemric, I., & Vujcic, B. (2002). Efficiency of Banks in Croatia: A DEA Approach. *Comparative Economic Studies*, 44(2–3), 169–193. <https://doi.org/10.1057/ces.2002.13>
- Joyce, O.-A. (2013). *Challenges of Deposit Mobilisation at Agricultural Development Bank* (Vol. 53, Issue 9). <https://doi.org/10.1017/CBO9781107415324.004>

- Kaffash, S., Kazemi, R., & Tajik, M. (2017). A directional semi-oriented radial DEA measure : An application on financial stability and the efficiency of banks. *Annals of Operations Research*. <https://doi.org/10.1007/s10479-017-2719-5>
- Kaitlin, ;, Smith, T. ;, & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review*, 1(3), 9.
- Kamarudin, F., Sufian, F., Loong, F. W., & Anwar, N. A. M. (2017). Assessing the domestic and foreign Islamic banks efficiency: Insights from selected Southeast Asian countries. *Future Business Journal*, 3(1), 33–46. <https://doi.org/10.1016/j.fbj.2017.01.005>
- Kartasheva, A. V., & Traskin, M. (2011). *Insurers ' Insolvency Prediction using Random Forest Classification Insurers ' Insolvency Prediction using Random Forest Classification*.
- Kim, K., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems With Applications*, 19, 125–132.
- Kılıçaslan, Y., Sickles, R. C., Kayış, A. A., & Gürel, Y. Ü. (2017). *Impact of ICT on the Productivity of the Firm: Evidence from Turkish Manufacturing*.
- Ko, M. (2004). *Using regression splines to assess the impact of information technology investments on productivity in the health care industry*. 43–63.
- Ko, M., & Kweku-Muata, O.-B. (2014). Reexamining the Impact of Information Technology Investments on Productivity Using Regression Tree and MARS-Based Analyses. *Advances in Research Methods for Information Systems Research, Integrated Series in Information Systems* 34. <https://doi.org/10.1007/978-1-4614-9463-8>
- Ko, M., & Osei-Bryson, K.-M. (2004). Using regression splines to assess the impact of information technology investments on productivity in the health care industry. *Information Systems Journal*, 14(1), 43–63. <https://doi.org/10.1111/j.13652575.2004.00160.x>
- Ko, M., & Osei-Bryson, K.-M. (2006). Analyzing the impact of information technology investments using regression and data mining techniques. *Journal of Enterprise Information Management*, 19(4), 403–417. <https://doi.org/10.1108/17410390610678322>
- Koellinger, P. (2005). *Why IT matters: An empirical study of e-business usage, innovation, and firm performance* (No. 495).
- Koivo, H. N. (2008). *Neural Networks : Basics using MATLAB Neural Network Toolbox*.
- Kotsiantis, S. B. (2007). Supervised Machine Learning : A Review of Classification Techniques. *Informatica*, 31, 249–268.
- Kriesel, D. (2005). *Inroduction to Neural Networks*. dkriesel.com.
- Krishnapuram, B., Carin, L., Figueiredo, A. T., & Hartemink, A. J. (2005). Sparse Multinomial Logistic Regression : Fast Algorithms and Generalization Bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6), 957–968.

- Kubat, M., Cholté, R., & Matwin, S. (1998). Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning*, 30, 195–215.
- Kudryavtseva, T., Rodionov, D., & Skhvediani, A. (2018). An empirical study of information technology clusters and regional economic growth in Russia. *IV International Scientific Conference*, 50, 1–11.
- Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of Diversity in Classifier Ensembles. *Machine Learning*, 51, 181–207.
- Kwon, H., & Lee, J. (2015). Two-stage production modeling of large U.S. banks: a DEANEural network approach. *Expert Systems with Application*.
<https://doi.org/10.1016/j.eswa.2015.04.062>
- Lam, M. (2004). Neural network techniques for financial performance prediction : integrating fundamental and technical analysis. *Decision Support Systems*, 37, 567–581.
[https://doi.org/10.1016/S0167-9236\(03\)00088-5](https://doi.org/10.1016/S0167-9236(03)00088-5)
- Lampe, H. W., & Hilgers, D. (2014). Trajectories of efficiency measurement: A bibliometric analysis of DEA and SFA. *European Journal of Operational Research*.
<https://doi.org/10.1016/j.ejor.2014.04.041>
- Lampe, H. W., & Hilgers, D. (2015). Trajectories of efficiency measurement: A bibliometric analysis of DEA and SFA. *European Journal of Operational Research*, 240(1), 1–21.
<https://doi.org/10.1016/j.ejor.2014.04.041>
- LaPlante, A. E., & Paradi, J. C. (2014). Evaluation of bank branch growth potential using data envelopment analysis. *The International Journal of Management Science*, 52, 23–41. <https://doi.org/10.1016/j.omega.2014.10.009>
- Le, J. (2018). *A Tour of The Top 10 Algorithms for Machine Learning Newbies*. Towards Data Science. <https://towardsdatascience.com/a-tour-of-the-top-10-algorithmsformachine-learning-newbies-dde4edffae11>
- Lee, S. (2010). Using data envelopment analysis and decision trees for efficiency analysis and recommendation of B2C controls. *Decision Support Systems*, 49(4), 486–497.
<https://doi.org/10.1016/j.dss.2010.06.002>
- Lefley, F. (2015). The Perception That ICT Projects Are Different. *The FAP Model and Its Application in the Appraisal of ICT Projects*, 21–22.
- Leung, L., & Zhang, R. (2016). Mapping ICT Use at Home and Telecommuting Practices: A Perspective from Work/Family Border Theory. *Telematics and Informatics*.
<https://doi.org/10.1016/j.tele.2016.06.001>
- Liaw, A., & Matthew, W. (2018). *Package “randomForest”* (4.6-14).
<https://doi.org/10.1023/A>
- Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(December), 18–22.
- Lim, T.-S., & Yu-Shan, S. (2000). A Comparison of Prediction Accuracy , Complexity , and Training Time of Thirty-Three Old and New Classification Algorithms. *Machine Learning*, 40(1992), 203–228.

- Livingston, F. (2005). Implementation of Breiman ' s Random Forest Machine Learning Algorithm. In *ECE591Q Machine Learning Journal Paper*.
- Luca, G. D. E., Riveccio, G., & Zuccolotto, P. (2010). Combining Random Forest and Copula Functions : A Heuristic Approach for Selecting Assets from a Financial Crisis Perspective. *Intelligent Systems in Accounting ,Finance and Management*, 109, 91–109. <https://doi.org/10.1002/isaf>
- Lusigi, A., & Ã, C. T. (1997). *Total Factor Productivity and the Effects of R & D in African Agriculture*. 9(4), 529–538.
- Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on A Prediction of P2P Network Loan Default Based on the Machine Learning LightGBM and XGboost Algorithms according to Different High Dimensional Data Clēaning *Electronic Commerce Research and Applications*. <https://doi.org/10.1016/j.elerap.2018.08.002>
- Madebo, S. A. (2013). *Factors Affecting Deposit Mobilization in Private Commercial Banks: the Case of Awash International Bank S.C. Addis Ababa, Ethiopia* [St. Mary's University College]. [http://repository.smuc.edu.et/bitstream/123456789/873/1/SISAY ASSEFA MADEBO.pdf](http://repository.smuc.edu.et/bitstream/123456789/873/1/SISAY_ASSEFA_MADEBO.pdf)
- Madjid, T., Mohammad, H. K., & Mohsen, J.-S. (2009). Information technology ' s impact on productivity in conventional power plants. *International Journal of Business Performance Management*, 11(3), 187–202.
- Mai, F., Tian, S., Lee, C., & Ma, L. (2018). Deep Learning Models for Bankruptcy Prediction using Textual Disclosure. *European Journal of Operational Research*. <https://doi.org/10.1016/j.ejor.2018.10.024>
- Majeed, M. T., & Ayub, T. (2018). Information and communication technology (ICT) and economic growth nexus: A comparative global analysis. *Pakistan Journal of Commerce and Social Sciences (PJCSS)*, 12(2).
- Maletić, R., Kreća, M., & Maletić, P. (2013). Application of DEA Methodology in Measuring Efficiency in the Banking Sector. *Economics of Agriculture*, 4, 843–855.
- Mashat, A. F., Fouad, M. M., Yu, P. S., & Gharib, T. . (2012). A Decision Tree Classification Model for University Admission System. *International Journal of Advanced Computer Science and Applications*, 3(10), 17–21.
- Mehrabiana, S. (2013). Using Non-Archimedean DEA Models for Classification of DMUs : A New Algorithm. *International Journal of Data Envelopment Analysis*, 1(4), 247–257.
- Mohd Zaini Abd, K. (2001). Comparative Bank Efficiency across Select ASEAN Countries. *ASEAN Economic Bulletin*, 18(3), 289–304. <https://doi.org/10.1355/AE18-3D>
- Mostafa, M. M. (2009). Modeling the efficiency of top Arab banks : A DEA – neural network approach. *Expert Systems With Applications*, 36(1), 309–320. <https://doi.org/10.1016/j.eswa.2007.09.001>
- Mousavi, M. M., Ouenniche, J., & Tone, K. (2019). A comparative analysis of two-stage distress prediction models. *Expert Systems with Applications*, 119, 322–341. <https://doi.org/10.1016/j.eswa.2018.10.053>

- Muhammad, N. K., Amin, M. F. Bin, Khokhar, I., Hassan, M. ul, & Ahmad, K. (2018). Efficiency Measurement of Islamic and Coventional Banks in Saudi Arabia : An Empirical and Comparactive Analysis. *Journal of Islamic Thought and Civilization of the International Islamic Universit Malaysia*, December.
- Mushtaq, S., & Siddiqui, D. A. (2017). Effect of interest rate on bank deposits: Evidences from Islamic and non-Islamic economies. *Future Business Journal*, 3(1), 1–8. <https://doi.org/10.1016/j.fbj.2017.01.002>
- Nahidul, S. M., Mohammed, I., Ali, J., & Wafik, H. M. A. (2019). Determinants of Deposit Mobilization of Private Commercial Banks : Evidence from Bangladesh. *International Journal of Business and Management Invention*, 8(10), 26–33.
- Nami, S., & Shajari, M. (2018). Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors. *Expert Systems With Applications*. <https://doi.org/10.1016/j.eswa.2018.06.011>
- Nand, kumar, & Archana, S. (2015). Measuring Technical and Scale Efficiency of Banks in India Using DEA. *IOSR Journal of Business and Management (IOSR-JBM)*, 17(1), 66–71. <https://doi.org/10.9790/487X-17126671>
- Nations, U. (2008). Measuring the impact of ICT use in business. The Case of Manufacturing in Thailand. *United Nations Conference on Trade and Development*.
- Navapan, K., Liu, J., & Sriboonchitta, S. (2017). Cost Efficiency of Top Thai Banks : A Comparison of Classical Stochastic Frontier with Efficiency Stochastic Frontier Models. *Thai Journal of Mathematics, Special Issue on Entropy in Econometrics*, 159–173.
- Navot, A. (2006). *On the Role of Feature Selection in Machine Learning* (Issue December). Hebrew University.
- Necmi K, A. (2006). Productivity Analysis in the Service Sector with Data Envelopment Analysis. In *Analysis* (Third Edit). N K Avkiran UQ Business School The University of Queensland QLD 4072, Australia.
- Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Expert Systems with Applications Application of data mining techniques in customer relationship management : A literature review and classification. *Expert Systems With Applications*, 36(2), 2592–2602. <https://doi.org/10.1016/j.eswa.2008.02.021>
- Niebel, T., & Mannheim, Z. E. W. (2014). ICT and Economic Growth - Comparing Developing , Emerging and Developed Countries. *IARIW 33rd General Conference*.
- Ogunde, A. ., & Ajibade, D. . (2014). A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm. *Journal of Computer Science and Information Technology*, 2(1), 21–46.
- Olaru, C., & Wehenkel, L. (2003). A complete fuzzy decision tree technique. *Fuzzy Sets and Systems*, 138, 221–254. [https://doi.org/10.1016/S0165-0114\(03\)00089-7](https://doi.org/10.1016/S0165-0114(03)00089-7)
- Omary, Z., & Mtenzi, F. (2010). Machine Learning Approach to Identifying the Dataset Threshold for the Performance Estimators in Supervised. *International Journal for Infonomics (IJI)*, 3(3). <https://doi.org/10.20533/iji.1742.4712.2010.0034>
- Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., &

- Akinjobi, J. (2017). Supervised Machine Learning Algorithms : Classification and Comparison. *International Journal of Computer Trends and Technology(IJCTT)*, 48(3). <https://doi.org/10.14445/22312803/IJCTT-V48P126>
- Ostadi, H., & Sarlak, A. (2014). Effective factors on the absorption of bank deposits in order to increase the relative share of Isfahan Sepah Bank. *International Journal of Academic Research in Economics and Management Sciences*, 3(4), 139–149. <https://doi.org/10.6007/ijarems/v3-i4/1112>
- P.R.A. Oeij, M.P. De Looze, K. Ten Have, J.W. Van Rhijn, L. F. M. K. (2011). *Developing the organization ' s productivity strategy in various sectors of industry*. <https://doi.org/10.1108/17410401211187525>
- Paço, C. M. L., & Pèrez, J. M. C. (2015). Assessing the impact of Information and Communication Technologies on the Portuguese hotel sector : an exploratory analysis with Data Envelopment Analysis. *Tourism & Management Studies*, 11(1), 35–43.
- Pandya, R., & Pandya, J. (2015). C5 . 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications*, 117(16), 18–21.
- Paradi, J. ., Vela, S. ., & Zhu, H. (2010). Bank Branch Operational Studies Using DEA. *International Series in Operations Research & Management Science* 266.
- Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2014a). Predicting stock and stock price index movement using Trend Deterministic Data Preparation and Machine Learning Techniques. *Expert Systems With Applications*, August. <https://doi.org/10.1016/j.eswa.2014.07.040>
- Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2014b). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Application*, October. <https://doi.org/10.1016/j.eswa.2014.10.031>
- Pesa, E. M. O., & Muturi, W. (2015). Factors Affecting Deposit Mobilization By Bank Agents in Kenya : a Case of National Bank of Kenya , Kisii County. *International Journal of Economics, Commerce and Management*, III(6), 1545–1557.
- Point, T. P. (I) P. L. (2016). *R Programming Tutorialspoint SimplyEasyLearning*.
- Portas, J., & AbouRizk, S. (1997). Neural Network Model for Estimating Construction Productivity. *Journal of Construction Engineering and Management*, December, 399–410.
- Premachandra, I. M., Bhabra, G. S., & Sueyoshi, T. (2009). DEA as a tool for bankruptcy assessment: A comparative study with logistic regression technique. *European Journal of Operational Research*, 193(2), 412–424. <https://doi.org/10.1016/j.ejor.2007.11.036>
- Qiu, M., Song, Y., & Akagi, F. (2016). Application of artificial neural network for the prediction of stock market returns : The case of the Japanese stock market. *Chaos, Solitons and Fractals Nonlinear Science, and Nonequilibrium and Complex Phenomena*, 85, 1–7. <https://doi.org/10.1016/j.chaos.2016.01.004>
- Razavi, S. H., Amoozad, H., Zavadskas, E. K., & Hashemi, S. S. (2013). A Fuzzy Data Envelopment Analysis Approach based on Parametric Programming. 8(4), 594–607.

- Ream, R. K., & Rumberger, R. W. (2008). Student Engagement, Peer Social Capital, and School Dropout Among Mexican American and Non-Latino White Students. *Sociology of Education*, 81(2), 109–139. <https://doi.org/10.1177/003804070808100201>
- Rocha, B. C. da, & Júnior, R. T. de S. (2010). Identifying Bank Frauds Using CRISP-DM AND Decisions Trees. *International Journal of Computer Science & Information Technology (IJCSIT)*, 2(5), 162–169.
- Roghieh, G., Saeed, M., & Sang-Yong, T. L. (2004). ICT and Productivity of the Manufacturing Industries in Iran. *The Electronic Journal on Information Systems in Developing Countries*, 1–26.
- Ross, J. . Q. (1994). *C4 . 5 : Programs for Machine Learning* (A. Segre (ed.); Review Edi, Vol. 240). 1994 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.
- Roy, A. G., & Urolagin, S. (2019). Credit Risk Assessment Using Decision Tree and Support Vector Machine Based Data Analytics. *Advances in Science, Technology & Innovation (IEREK Interdisciplinary Series for Sustainable Development)*, 79–84. https://doi.org/10.1007/978-3-030-01662-3_10
- Saher, T., Malik, M. A. S., Ullah, S., & Ullah, A. (2019). Measuring Firm and Sector Efficiency in Pakistan: An Application of Data Envelopment Analysis. *Studies in Business and Economics*, 14(14), 239–257. <https://doi.org/10.2478/sbe-2019-0057>
- Sahoo, S. (2014). Financial Intermediation and Growth : Bank-Based versus Market-Based Systems. *The Journal of Applied Economic Research*, 2, 93–114. <https://doi.org/10.1177/0973801013519998>
- Sakouvogui, K. (2019). Banks performance evaluation: A hybrid DEA-SVM- The case of U.S. agricultural banks Kekoura Sakouvogui. *Accounting*, 5, 107–120. <https://doi.org/10.5267/j.ac.2018.09.002>
- Santos, S. P., Belton, V., Howick, S., & Pilkington, M. (2017). Technological Forecasting & Social Change Measuring organisational performance using a mix of OR methods. *Technological Forecasting & Social Change*, January, 0–1. <https://doi.org/10.1016/j.techfore.2017.07.028>
- Sarifuddin, S., Ismail, M. K., & Kumaran, V. V. (2015). Comparison of Banking Efficiency in the selected ASEAN Countries during the Global Financial Crisis. *Persidangan Kebangsaan Ekonomi Malaysia Ke-10*, 10(September), 286–293.
- Scornet, E. (2010). Consistency of Random Forests. *2010 Mathematics Subject Classification: 62G05, 62G20.*, 1–47.
- Shalev-Shwartz, S., & Ben-david, S. (2014). *Understanding Machine Learning : From Theory to Algorithms*. Cambridge University Press.
- Shamisi, M. H. Al, Assi, A. H., & Hejase, H. A. N. (2011). Using MATLAB to Develop Artificial Neural Network Models for Predicting Global Solar Radiation in Al Ain City – UAE. *Engineering Education and Research Using MATLAB*. <http://www.intechopen.com/books/engineering-education-and-research-usingmatlab/using-matlab-to-develop-artificial-neural-network-models-for-predictingglobalsolar-radiation-in-al>

- Shao, B. B. M., & Lin, W. T. (2001). Measuring the value of information technology in technical efficiency with stochastic production frontiers. *Information and Software Technology*, 43(7), 447–456. [https://doi.org/10.1016/S0950-5849\(01\)00150-1](https://doi.org/10.1016/S0950-5849(01)00150-1)
- Sharif, O., Hasan, M. Z., Kurniasari, F., Hermawan, A., & Gunardi, A. (2019). Productivity and efficiency analysis using DEA: Evidence from financial companies Listed in Bursa Malaysia. *Management Science Letters*, 9, 301–312. <https://doi.org/10.5267/j.msl.2018.11.010>
- Sheng, Y. P., & Mykytyn, Peter P., J. (2002). Information Technology Investment and Firm Performance : A Perspective of Data Quality. *Proceedings of the Seventh International Conference on Information Quality (ICIQ-02)*, 132–141.
- Shibu, D., & Ayekpam, I. C. (2018). A Study on Efficiency on Assam Gramin Vikash Bank Branches. *Indian Journal of Reserach*, 7(5), 115–118.
- Sigala, M. (2003). The information and communication technologies productivity impact on the UK hotel sector. *International Journal of Operations & Production*. <http://www.emeraldinsight.com/doi/abs/10.1108/01443570310496643>
- Silva, T. C., Tabak, B. M., Cajueiro, D. O., & Villas, Marina Dias, B. (2018). Adequacy of deterministic and parametric frontiers to analyze the efficiency of Indian commercial banks \$. *Physica A*. <https://doi.org/10.1016/j.physa.2018.04.100>
- Simm, J. (2016). *Robust Data Envelopment Analysis (DEA) for R Description*.
- Smola, A., & Vishwanathan, S. V. N. (2008). *Introduction to Machine Learning*. Cambridge University Press.
- Sousa, M. D. M., & Figueiredo, R. S. (2014). Credit Analysis Using Data Mining: Application in the case of Cedit Union. *JISTEM - Journal of Information Systems and Technology Management*, 11(2), 379–396. <https://doi.org/10.4301/S180717752014000200009>
- Sreedhara, B. M., Rao, M., & Mandal, S. (2018). Application of an evolutionary technique (PSO – SVM) and ANFIS in clear-water scour depth prediction around bridge piers. *Neural Computing and Applications*, 6. <https://doi.org/10.1007/s00521-018-3570-6>
- Sreekumar, S., & Mahapatra, S. S. (2011). Performance modeling of Indian business schools : a DEA-neural network approach. *Benchmarking: An International Journal*, 18(2), 221–239. <https://doi.org/10.1108/14635771111121685>
- Stanley, T. D., Doucouliagos, H., & Steel, P. (2018). Does ICT Generate Economic Growth? A Meta-Regression Analysis. *Journal of Economics Surveys*, 32(3), 705–726. <https://doi.org/10.1111/joes.12211>
- Steinke, J., & Etten, J. Van. (2017). Gamification of farmer-participatory priority setting in plant breeding : Design and validation of “ AgroDuos .” *Journal of Crop Improvement*, 0(0), 1–23. <https://doi.org/10.1080/15427528.2017.1303801>
- Stewart, C., Matousek, R., & Nguyen, T. N. (2015). Efficiency in the Vietnamese banking system: a DEA double bootstrap approach. *Research in International Business and Finance*. <https://doi.org/10.1016/j.ribaf.2015.09.006>

- Sufian, F., Kamarudin, F., & Nassir, A. md. (2016). Determinants of efficiency in the malaysian banking sector: Does bank origins matter? *Intellectual Economics*, 10(1), 38–54. <https://doi.org/10.1016/j.intele.2016.04.002>
- Syverson, C. (2011). What Determines Productivity? *Journal of Economic Literature*, 49(2), 326–365. <https://doi.org/10.1257/jel.49.2.326>
- Tamura, S., & Tateishi, M. (1997). Capabilities of a four-layered feedforward neural network: four layers versus three. *Proceedings of the IEEE Transactions on Neural Networks*, 251–255. <https://doi.org/10.1111/j.1467-8535.2011.01259.x>
- Tanaka, K., Kinkyō, T., & Hamori, S. (2016). Random forests-based early warning system for bank failures. *Economics Letters*. <https://doi.org/10.1016/j.econlet.2016.09.024>
- Tang, D. (2016). *Random Forest* (Issue May, pp. 1–9).
- Tavana, M., Fallahpour, A., Di Caprio, D., & Santos-Arteaga, F. J. (2016). A hybrid intelligent fuzzy predictive model with simulation for supplier evaluation and selection. *Expert Systems with Applications*, 61, 129–144. <https://doi.org/10.1016/j.eswa.2016.05.027>
- Tavana, M., Izadikhah, M., Caprio, D. Di, & Saen, R. F. (2017). A New Dynamic Range Directional Measure for Two-Stage Data Envelopment Analysis Models with Negative Data. *Computers & Industrial Engineering*. <https://doi.org/10.1016/j.cie.2017.11.024>
- Thanassoulis, E., Portela, M. C. S., & Despi, O. (2008). The Mathematical Programming Approach to Efficiency Analysis. In *DEA – The Mathematical Programming Approach to Efficiency Analysis* (pp. 1–161). Oxford University Press.
- Titko, J., Stankevičienė, J., & Lāce, N. (2014). Measuring bank efficiency: DEA application. *Technological and Economic Development of Economy*, 20(4), 739–757. <https://doi.org/10.3846/20294913.2014.984255>
- Toloo, M., Zandi, A., & Emrouznejad, A. (2015). Evaluation efficiency of large-scale data set with negative data : an artificial neural network approach. *Journal of Supercomputing*. <https://doi.org/10.1007/s11227-015-1387-y>
- Tra, N. T., Minh, L. Q., Phuc, C., Khoa, T., & Thanh, N. P. (2018). Incorporating Risk into Technical Efficiency via a Semiparametric Analysis : The Case of ASEAN Banks. *VNU Journal of Science: Economics and Business*, 34(2), 54–64.
- Tsai, C., & Wu, J. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Application*, 34, 2639–2649. <https://doi.org/10.1016/j.eswa.2007.05.019>
- Tsai, M., Lin, S., Cheng, C., & Lin, Y. (2009). The consumer loan default predicting model – An application of DEA – DA and neural network. *Expert Systems With Applications*, 36(9), 11682–11690. <https://doi.org/10.1016/j.eswa.2009.03.009>
- Turhani, A., & Hoda, H. (2016). The Determinative Factors of Deposits Behavior in Banking System in Albania (Jan 2005 – Dec 2014). *Academic Journal of Interdisciplinary Studies*, 5(2), 246–256. <https://doi.org/10.5901/ajis.2016.v5n2p246>
- Tzeng, K. S. G. (2014). A decision rule-based soft computing model for supporting financial performance improvement of the banking industry. *Soft Computing*. <https://doi.org/10.1007/s00500-014-1413-7>

- Van Sang, H., Ha Nam, N., & Duc Nhan, N. (2016). A Novel Credit Scoring Prediction Model based on Feature Selection Approach and Parallel Random Forest. *Indian Journal of Science and Technology*, 9(20). <https://doi.org/10.17485/ijst/2016/v9i20/92299>
- Vidyarthi, H. (2018). Dynamics of intellectual capitals and bank efficiency in India. *The Service Industries Journal*, 1–24. <https://doi.org/10.1080/02642069.2018.1435641>
- Vujičić, Tijana Matijević, T., & Zoran, Š. (2016). Comparative Analysis of Methods for Determining Number of Hidden Neurons in Artificial Neural Network. *Conference on Information and Intelligent Systems*, 219–223.
- Vuong, B. N., Tung, D. D., Giao, H. N. K., Dat, N. T., & Quan, T. N. (2020). Factors Affecting Savings Deposit Decision of Individual Customers: Empirical Evidence from Vietnamese Commercial Banks. *The Journal of Asian Finance, Economics and Business*, 7(7), 293–302. <https://doi.org/10.13106/jafeb.2020.vol7.no7.293>
- Wang, C. H., Gopal, R. D., & Zionts, S. (1997). Use of data envelopment analysis in assessing Information technology impact on firm performance. *Annals of Operations Research*, 73, 191–213. <https://doi.org/10.1023/A:1018977111455>
- Wang, C. H., Gopal, R. D., & Zionts, S. (1997). *Use of Data Envelopment Analysis in assessing Information Technology impact on firm performance*. 73, 191–213.
- Wang, C., Luu, Q., & Nguyen, T. (2019). Assessing Bank Performance Using Dynamic SBM Model. *Mathematics*. <https://doi.org/10.3390/math7010073>
- Wang, H., Wang, Y., Zhao, S., Wang, L., & An, H. (2018). The transnational comparative study on the potential risks and efficiency of commercial banks based on the weightlimited DEA model. *China Finance Review International*. <https://doi.org/10.1108/CFRI06-2017-0126>
- Wanke, P., Barros, C. P., & Emrouznejad, A. (2015). Assessing Productive Efficiency of Banks Using Integrated Fuzzy-DEA and Bootstrapping: A Case of Mozambican Banks. *European Journal of Operational Research*. <https://doi.org/10.1016/j.ejor.2015.10.018>
- Wanke, P., Barros, C. P., & Emrouznejad, A. (2016). A Comparison Between Stochastic DEA and Fuzzy DEA Approaches : Revisiting Efficiency in Angolan Banks. *RAIRO Operations Research*, 52, 285–303.
- Wanke, P., Kalam Azad, M. A., Barros, C. P., & Hadi-Vencheh, A. (2016). Predicting performance in ASEAN banks: an integrated fuzzy MCDM–neural network approach. *Expert Systems*, 33(3), 213–229. <https://doi.org/10.1111/exsy.12144>
- West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, 27, 1131–1152.
- Wickramasinghe, V., & Karunasekara, M. (2016). *Perceptual differences of enterprise resource planning systems between management and operational end-users*. 3001(August). <https://doi.org/10.1080/0144929X.2010.528027>
- Wolf, S. (2001). Determinants and Impact of ICT Use for African SMEs Implications for Rural South Africa. *DESG-IESG Annual Conference 2001*.
- Wong, J., & Dow, K. E. (2011). The Effects of Investments in Information Technology on

- Firm Performance : An Investor Perspective. *Journal of Information Technology Research*, 4(September(3)), 1–13. <https://doi.org/10.4018/jitr.2011070101>
- Wu, D. (2006). Detecting information technology impact on firm performance using DEA and decision tree. *Int. J. Information Technology and Management*, 5, 162–174.
- Wu, D. (2009). Supplier selection : A hybrid model using DEA , decision tree and neural network. *Expert Systems With Applications*, 36(5), 9105–9112. <https://doi.org/10.1016/j.eswa.2008.12.039>
- Wu, D. D., Yang, Z., & Liang, L. (2006). Efficiency analysis of cross-region bank branches using fuzzy data envelopment analysis. *Applied Mathematics and Computation*, 181, 271–281. <https://doi.org/10.1016/j.amc.2006.01.037>
- Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., Mclachlan, G. J., Ng, A., Liu, B., Yu, P. S., Michael, Z. Z., David, S., & Dan, J. H. (2008). Top 10 algorithms in Data mining. *Knowledge and Information Systems*, 1–37. <https://doi.org/10.1007/s10115-007-0114-2>
- Yang, C., Zou, Y., Lai, P., & Jiang, N. (2015). Data mining-based methods for fault isolation with validated FMEA model ranking Mean Time between Failure. *Applied Intelligence*, 43, 913–923. <https://doi.org/10.1007/s10489-015-0674-x>
- Yao, Y., Viswanath, B., Xiao, Z., Zheng, H., Wang, B., & Zhao, B. Y. (2017). Complexity vs. Performance: Empirical analysis of machine learning as a service. *Proceedings of the ACM SIGCOMM Internet Measurement Conference, IMC, Part F1319*(119), 384–397. <https://doi.org/10.1145/3131365.3131372>
- Yeo, B., & Grant, D. (2018). Predicting service industry performance using decision tree analysis. *International Journal of Information Management*, 38(1), 288–300. <https://doi.org/10.1016/j.ijinfomgt.2017.10.002>
- Yousaf, H. (2016). *Analysing which factors are of influence in predicting the employee turnover*. Analysing which factors are of influence in predicting the employee turnover Research Paper Business Analytics HMN Yousaf Supervised by Dr. Sandjai Bhulai Vrije Universiteit Amsterdam.
- Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with. *J. R. Statist. Soc. B*, 68(1), 49–67.
- Zamar, D., McNeney, B., & Graham, J. (2007). elrm: Software implementing exact-like inference for logistic regression models. *Journal of Statistical Software*, 21(3), 1–18. <https://doi.org/10.18637/jss.v021.i03>
- Zha, Y., Liang, N., Wu, M., & Bian, Y. (2016). Efficiency evaluation of banks in China : A dynamic two-stage slacks-based measure approach. *Omega*, 60, 60–72. <https://doi.org/10.1016/j.omega.2014.12.008>
- Zheng, Z., Padmanabhan, B., & Zheng, H. (2004). A DEA Approach for Model Combination. *KDD'04, August 22–25, 2004. Research Track Poster*, 755–760.

DM U	No. of Employee s	% Performing Loans	IT budget(GH¢)	Fixed Asset(GH¢)	Deposit(GH¢)	Profit(GH¢)
1	7	83.7537	60897.3621	1454093.72	72757137	316871.7
2	7	82.69211	13859.0755	1458937.49	17605016	4372073
3	7	83.11167	34710.1547	58751.02	10256530	1939064.7
4	7	82.54663	23696.8498	105086.5	68830240	4579780.7
5	7	81.09985	47438.3618	1768645.92	30798389	778894.6
6	7	83.27801	27897.1118	231642.89	52158533	1098903.5
7	7	84.39781	29038.052	886766.52	79999706	2169496.8
8	7	84.7925	66112.7017	610520.6	12158202	4546746.5
9	7	84.38961	25474.0981	96518.98	23429068	2726352.2
10	7	83.91438	19228.4336	591265.4	66972317	1608772.5
11	7	83.14883	90718.4762	983547.47	44330728	1412219.6
12	7	84.21455	63402.9388	899117.96	5509682	1999396
13	7	82.71137	24820.0458	1113390.03	63062442	3485935.7
14	7	82.12436	82198.1167	895869.7	42907341	4601286.4
15	10	81.28045	66849.9966	378360.87	47005611	2937952.7
16	10	82.86733	41662.9971	1291477.35	56411163	178020.6
17	10	83.40505	13954.9107	157818.84	40767268	2197418.9
18	10	81.38958	8618.368	321805.62	73669458	1849928.7
19	10	84.21942	65122.7957	74096.2	4045097	356569.6
20	10	82.80137	22501.4046	1195705.25	26949380	1249464.5
21	8	83.44796	74708.9867	1648052.42	30161464	4420331.2
22	8	82.39093	71708.5146	934887.48	34945738	1202004.4
23	8	81.89895	51230.8662	1312692.06	72541502	275771.1
24	8	81.72461	72807.1954	145703.12	33410697	1746629.5

KNUST



Zhiyu, W. (2016). *Fast Estimation of Multinomial Logit Models* : 75(3).
<https://doi.org/10.18637/jss.v075.i03>

Zhou, X., Xu, Z., Chai, J., Yao, L., Wang, S., & Lev, B. (2018). Efficiency evaluation for banking systems under uncertainty: A multi-period three-stage DEA model. *Omega*.
<https://doi.org/10.1016/j.omega.2018.05.012>

APPENDICES

Appendix A: The DEA Inputs and Outputs data for the 444 Banks DMUs

25	8	82.24098	57001.2629	1343246.94	4555214	214198.7
26	8	81.9301	56761.132	617310.01	1917982	2766706.5
27	8	83.39029	61384.8137	1108729.48	51759575	881929.5
28	8	83.4618	75610.5927	731615.58	85077214	2686259.8
29	8	83.7893	41301.2897	1223918.68	51156498	3748499.9
30	8	81.61138	16147.4105	1241505.32	44833724	1984384.1
31	8	84.66789	64261.9116	651393.03	76198639	1725473.9
32	8	82.91391	74086.2025	862395.61	53541075	4735426.5
33	8	82.97651	37068.7217	364839.11	83902102	3064981.4
34	8	84.00542	35709.9788	659382.86	59601194	337238.8
35	7	84.48814	89137.2205	1217831.91	33364809	1792806.8
36	7	82.9628	22715.4057	456158.54	60340108	497863.4
37	7	83.9244	65887.3364	134009.64	10247869	2698873.5
38	7	83.34419	87880.7609	1087680.88	49789686	3924384.6
39	7	84.73585	58948.1106	1563957.17	39143763	4601904.2
40	7	83.24943	530.0299	1721702.77	33800644	1364603.2
41	7	83.81039	61867.3149	2404364.27	61626273	60831.48
42	7	84.21065	27642.0198	1615440.56	60231082	383397.99
43	7	85.38675	43620.7308	1211423.16	86922361	94964.24
44	7	85.57497	65135.1257	317603.59	38406087	23864.06
45	7	85.73434	134827.1346	2306880.07	31366089	154732.67
46	7	83.97825	39624.0834	2312024.67	56303255	157540.13
47	7	84.61915	27492.8512	2113253.82	70861614	22598.56
48	7	86.72497	125135.3219	1812950.03	61112071	330762.65
49	7	85.71362	132067.1273	10488.58	36145271	262307.63
50	7	85.98953	120439.421	1138134.67	79288312	153185.81
51	7	84.39833	118992.3979	1884590.82	39256592	374745.47
52	7	85.70841	137779.1401	454888.96	49350727	87500.86
53	7	83.54879	127492.9429	1310507.91	42026118	263872.76
54	7	85.46946	109222.5359	1441672.79	10186527	10416.49
55	8	84.33161	131057.0528	2406397.4	15222349	254451.9
56	8	84.84177	107542.4125	2230702.51	67899842	376535.7
57	8	84.7416	114753.5643	486385.17	72277525	231460.49
58	8	86.83833	52398.3301	1945943.36	38314238	413824.09
59	8	85.13257	81484.6853	85604.18	33186431	32792.78
60	8	84.0898	52190.2872	1212942.5	57899684	141870.95
61	8	83.98471	29978.277	2175448.62	54740574	427959.06

KNUST



64	8	84.59234	112863.8048	2467311.67	39364242	224161.31
65	8	84.71331	45121.2452	214377.29	2414738	376569.97
66	8	84.27656	92631.6484	1456254.17	73009061	129133.45
67	8	84.73299	61014.6942	568709.68	20918328	243528.33
68	8	84.18429	109359.861	2306232.74	15808106	179951.77
69	8	83.39657	87069.6778	233394.13	32962703	50624.38
70	8	85.69877	125409.982	2105899.64	9106922	214176.49
71	8	84.73139	1703.001	2227485.17	38741653	441839.07
72	8	83.76835	72075.1806	46138.3	75065899	93423.12
73	8	84.70759	6528.152	1811396.71	5944887	47912.09
74	10	85.39766	137866.8884	601675.05	21017254	411043.51
75	6	84.65154	85611.8689	1868360.33	17369357	377752.15
76	10	84.4557	27052.307	1046261.6	12494808	321648.99
77	10	85.14075	101813.8507	1473196.92	51589349	279664.51
78	10	84.60654	41963.3303	2250464.82	23130577	454686.37
79	10	84.41352	102690.5301	2980133.23	69203635	186247.24
80	10	85.89807	88988.9458	390943.05	56341724	327418.36
81	6	85.52212	127335.2224	2426354.68	44078449	299258.46
82	10	85.46361	102907.4984	2058839.68	70993344	138064.99
83	10	84.82036	102690.5301	2980133.23	69203635	186247.24
84	10	85.89807	88988.9458	390943.05	56341724	327418.36
85	6	85.52212	127335.2224	2426354.68	44078449	299258.46
86	10	85.46361	102907.4984	2058839.68	70993344	138064.99
87	10	83.42265	71745.8199	2294327.81	85805284	440015.77
88	10	84.49259	433.6275	1597652.84	11242352	108958.59
89	10	86.42369	41508.8025	98280.42	10034482	311570.81
90	7	85.19818	137110.8555	1933128	47475678	61259.5
91	7	84.32041	6736.4367	868209.5	84691512	51813.97
92	7	85.07102	137866.7712	2451903.63	4680542	40803.27
93	7	84.75843	104523.8738	102616.68	64848764	279102.96
94	7	84.68759	63633.6001	1776428.59	62553116	409229.01
95	7	83.61287	93886.5298	1163921.8	38618660	471542.49
96	7	85.11578	119991.1816	1280388.63	42924666	360773.33
97	7	84.58379	146730.6044	496686.5	71278156	401675.41
98	7	84.45135	113908.6877	739820.95	50491550	320079.93
99	7	85.49803	81657.5728	1544049.95	22010521	73544
100	7	83.63567	43702.7095	2424035.2	8549690	322669.72
101	7	84.02338	25250.4116	1062570.62	57244312	437573.05

KNUST



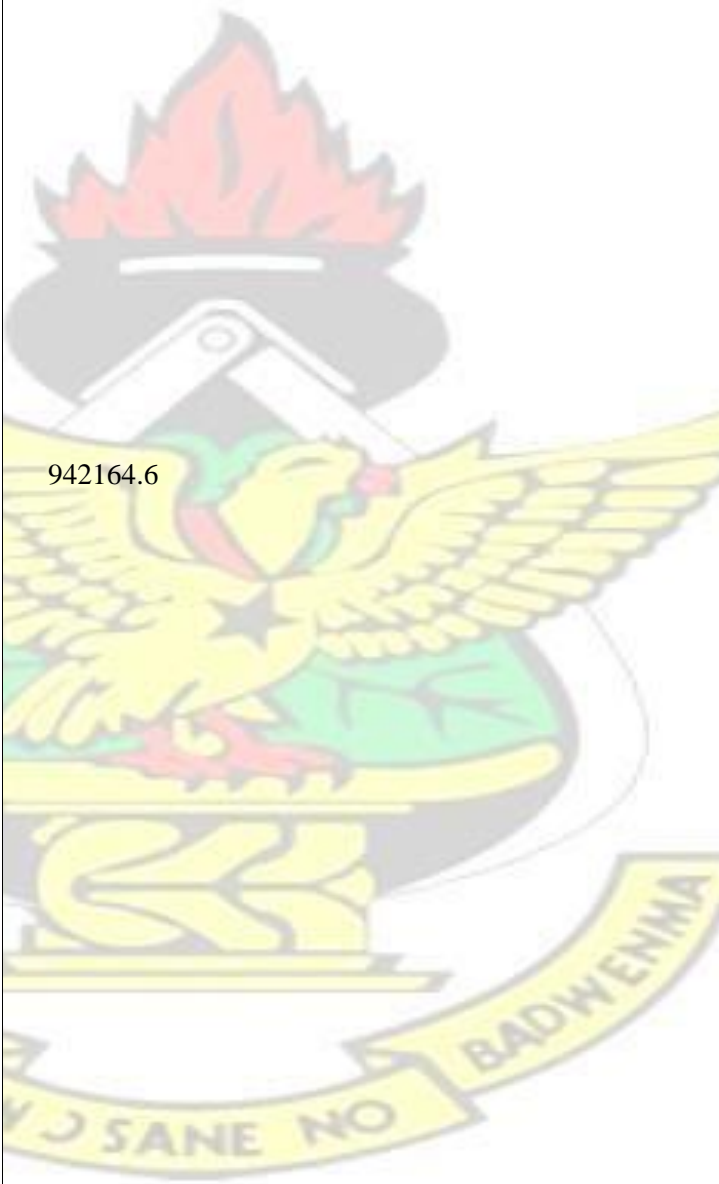
102	7	86.77478	118550.8003	1894949.44	5763210	422830.69
103	7	86.02005	34931.6889	2007830.7	32830717	451720.6
104	7	85.57658	111870.2223	2301530.17	14895089	235689.32
105	7	83.84095	30768.1097	542467.36	27911303	37370.9
106	7	85.46901	129297.9651	438451.89	47314637	238965.23
107	7	82.57352	90000.2404	1567939.23	71753702	186826.77
108	7	85.77748	20078.8769	2136184.04	79250751	53221.74
109	7	84.03513	1876.2385	1370517.52	66082441	153566.76
110	7	85.70363	60363.7279	225139.13	40460697	445369.2
111	7	86.00411	8589.9743	299801.69	1684319	123565.41
112	7	86.2237	113514.7764	1505380.23	53877282	470794.2
113	7	82.23805	58950.2078	1444318.8	83413731	306483.57
114	7	85.00876	81951.1983	941272.87	60223089	358067.28
115	7	86.0707	130391.9549	1961424.09	26043899	133689.59
116	7	86.03892	89617.2365	289939.02	25290064	291611.03
117	7	84.68415	119371.8028	1239981.13	60599844	458001.55
118	8	98.36131	321077.1	4105600.5	35013412	10510990.2
119	8	99.46096	2537551.9	1731176.1	35947326	1086271
120	8	97.28851	751690.7	6019263.6	72250087	7074081.1
121	8	98.47515	2133281.4	3232716.4	101952375	648978
122	8	98.88848	1292845.1	386082.3	9398731	6917074.8
123	8	96.0508	1653019.4	1595165.6	104024958	10945686
124	8	99.05209	2227578	4154594.8	142069489	5077987.6
125	8	97.52781	544934.5	5768070.1	6779230	138387.3
126	8	98.53376	2457982.3	6496063.6	34941322	3499332.3
127	8	99.01635	787993.8	3168993.9	7758351	7446669.9
128	8	97.65014	1767798.2	7265533.9	77888580	10342194.4
129	8	99.43064	2588620.9	1026130.1	128751589	3446815.6
130	7	98.25481	1209918.6	7371631.1	135912822	2646443
131	7	98	762624.1	813814.8	57914100	5252397.1
132	7	99.61498	501231.2	6283134.3	73438281	3076513.7
133	7	98.33	1072597.4	4577787.3	92001282	6874788.1
134	7	98.07267	768843	3373742.6	99321171	9196754
135	7	98.70586	1708760.9	5630683.7	43533557	8744041.1
136	7	97.05627	651801.4	6135158.9	7747081	8218472
137	7	98.23505	1224650.8	6801755.9	56694183	8881535

KNUST



14 0	799.9095 6	1562372.5	2171086.4	94516973	7034537. 4
14 1	798.4551 4	843531.6	2359478	82212382	3279579. 4
14 2	798.5443 4	1585753.4	6513016.1	13635962 1	8404108. 4
14 3	799.5776 4	2512064.6	2664817	42779358	7687387. 9
14 4	799.2556 3	481919.7	4295186	91371505	3223650. 5
14 5	798.8579 4	380896.6	176459.2	78563706	8539613. 5

14 6	798.2141 3	1474280	5880578.8	12250198 0	1800166. 6
14 7	799.3602	1753489.4	4149595.9	11044626 2	7668729. 2
14 8	798.1459 7	1770096.9	1615703.2	15360839 5	10568444 .9
14 9	798.1762 4	979183	4546131.3	37138741	629038.4
15 0	798.3329 4	663485.8	126533.9	11376335 0	8598456. 4
15 1	797.0015 2	2085136.4	961278	64878503	4058873. 6
15 2	782.0002 6	1108931.94	2506457.1 65	21362641. 8	5206934. 08
15 3	784.8554 6	326481.57	3766395.7 82	52237080. 3	5072298. 79
15 4	784.1622 3	751980.4	110919.21	239959.5	1846667. 4
15 5	782.3900 6	130032.52	4182459.5 37	10119667. 3	3578507. 92
15 6	783.1142 8	242556.34	3817606.2 02	43359028. 6	75559.23 8
15 7	784.4780 4	299437.14	3002470.8 55	33934353. 2	1081354. 51
15 8	783.6397 5	940722.56	1651442.1 91	40815109. 1	140021.0 86

159	784.33198	 <p>942164.6</p>	3597834.887	8979527.4	2405554.54
160	782.88102	1222675.04	4404799.944	34235253.7	2129945.4
161	782.28032	398132.84	4014768.378	32452348.4	3481403.73
162	781.52163	642158.85	302468.901	16189601.7	2297697.06
163	784.50359	342295.36	4210096.472	29084457.8	3358087.66
164	782.18409	612146.81	4138228.796	356886.1	3356740.01
165	785.36291	346850.82	2136418.364	26767732.3	3903962.9

16 6	7 6	83.8685 6	348692.61	3809241.0 96	52948924. 8	3235949. 68
16 7	7 4	84.1336 4	585061.92	257617.49	47837720. 5	5357837. 62
16 8	7 9	82.6606 9	183142.26	2735387.3 67	9297909. 8	3114191. 44
16 9	7 2	82.9925 2	616914.07	4327.485	37434551. 7	151150.5 15
17 0	7 8	84.4726 8	413012.14	3091436.36	41203646. 5	4283152. 2

17 1	7 9	82.802 9	449039.2	3833892.6 39	52844924 .9	4110030.6 3
17 2	7 5	82.013 5	891619.7	2206469.4 13	10045253 .9	3730017.7 3
17 3	8 33	83.210 33	252109.17	4194378.5 98	48251213 .6	3010627.3 4
17 4	8 17	83.167 17	139858.09	4264789.4 12	3825606. 6	7424.127
17 5	8 79	81.178 79	164231.81	4328487.0 57	2508880. 7	4264805.3 8
17 6	8 34	81.176 34	342981.24	242238.615	47507489 .5	2225206.3 3
17 7	8 98	82.751 98	599384.12	2837806.6	37492716 .4	3997277.8
17 8	8 2	85.0789 2	18187.86	1179946.1 05	50748985. 2	2983732. 24
17 9	8 7	83.4482 7	185988.24	3556427.5 53	30686046	2620285. 27
18 0	8 2	81.9460 2	506371.18	509307.737	18487116. 6	1726386. 71
18 1	8 8	84.9081 8	629024.6	88138.47	48886883. 6	113912.3 52
18 2	8 7	82.9676 7	779307.96	639475.288	5986640.1	4115126. 49
18 3	8 6	83.8923 6	257910.43	2490002.9 43	32447078	2996130. 46
18 4	8 4	81.9891 4	665520.5	2495753.2 19	50423673	1045263. 87
18 5	8 3	82.5017 3	744796.73	3776270.1 59	33095830. 8	5165246. 79
18 6	8 2	82.0144 2	284858.94	1840045.1 88	43680874. 4	1464142. 17
18 7	8 2	82.1576 2	809251.09	109092.966	52524259. 6	557581.9 72
18 8	8 6	84.0623 6	504028.44	1127782.1 43	47865636. 7	5135843. 49
18 9	8 3	83.2653 3	13260.05	2146417.36	40086973. 3	1382960. 05
19 0	8 1	81.5087 1	835344.57	2354995.9 84	18127613. 6	529845.9 59

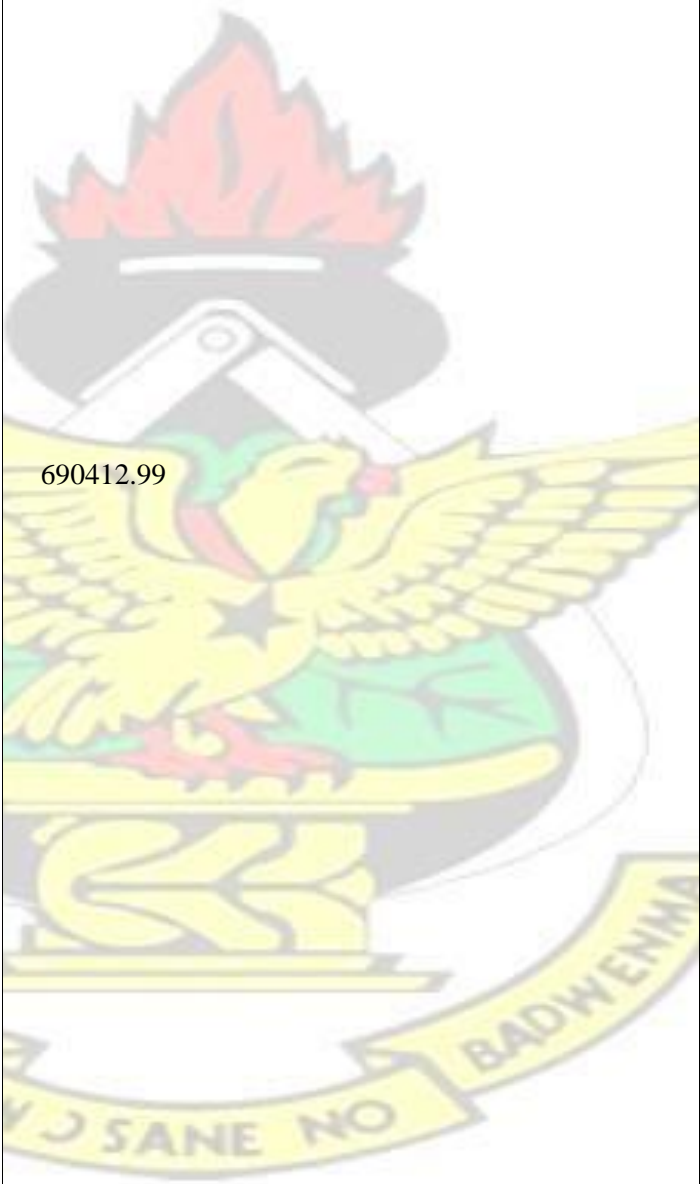
19 1	8	84.2615 1	591802.54	1364560.8 79	46215888. 5	1003650. 35
19 2	9	82.7230 7	377390.48	4361029.9 79	23438558. 2	3779573
19 3	9	82.7658	912260.15	1352236.7 88	45585888. 4	3855753. 05
19 4	9	83.4331 8	284462.72	4362167.1 28	29158933. 9	759982.9 66
19 5	9	83.2505 9	520732.5	972783.627	23366360. 5	911957.2 24
19 6	9	82.2216 6	202504.39	1568925.3 59	3493231.6	3372627. 16
19 7	9	82.2452 2	379526.19	498498.801	52206594. 9	2589170. 42

19 8	9	83.3073 7	914371.95	1873206.1 84	37833570. 3	1599196. 53
19 9	9	82.5892 4	812478.77	4212139.3 83	22130239. 5	1555236. 69
20 0	9	83.1944 6	511622.81	3449691.5 39	106844.8	3268024. 41
20 1	9	84.2021 8	873907.98	691358.26	35342167. 7	328620.6 15
20 2	9	81.6963 4	697783.73	1312174.3 36	20997534. 4	102770.4 79
20 3	7	83.3893 9	920241.04	2339441.3 47	36212390. 1	1349861. 49
20 4	7	83.7592 9	1002644.98	2263893.5 44	53658663. 7	4426091. 6
20 5	7	82.2098 9	1197804.19	3550038.3 04	34962145. 5	3380922. 03
20 6	7	81.8764 1	632321.45	2599828.9 17	30217487. 9	1814431. 76
20 7	7	81.5314 6	131873.99	4006209.9 53	3740852.1	2410676. 62
20 8	7	82.4320 8	819794.69	2158832.1 63	10874561. 6	3406117. 11
20 9	7	83.4572 5	276370.12	2884354.79	10300981	1846963. 81
21 0	7	85.5392 6	124126.24	682878.561	26423785. 7	3764212. 24
21 1	7	82.8574 4	438207.62	2358484.7 28	36034865. 5	4711171. 22
21 2	7	83.6501 8	1149509.81	1578214.2 98	36312850. 8	73941.14 3
21 3	7	82.3529 1	265045.14	2150124.3 31	40063069. 9	1330385. 29
21 4	7	83.0033 1	1048314.37	80128.112	16878809. 5	5258928. 44

215	7	81.35924	520558.49	1242748.596	9417916.2	5202649.24
-----	---	----------	-----------	-------------	-----------	------------

21 6	7	82.6245 1	838590.5	788708.032	10478538. 1	4456794. 37
21 7	7	82.5881 9	482478.5	1851866.83	45443835. 6	2447043. 69
21 8	7	83.7637	841223.56	3920988.8 93	28113694	2094294. 89
21 9	7	82.1520 8	123279.46	2460541.2 03	172946.5	4996613
22 0	7	82.9583 4	98276.5	2774433.4 51	25529915. 8	3435995. 46
22 1	7	81.9534 4	85753.41	3841123.9 84	9752038.2	1772478. 39
22 2	7	83.2766 9	449467.68	1113997.3 55	30401404. 5	297482.6 12

22 3	7	82.3587 7	1150160.89	234618.706	52839977. 7	2223829. 49
22 4	7	84.6407 4	387618.89	3028238.4 75	27477685	4064979. 71
22 5	7	84.2799 2	261949.13	2606305.1 96	37363993. 8	962988.6 55
22 6	7	83.5457 3	90203.64	17982.031	6080739.5	2018319. 18
22 7	7	81.4545 6	474819.59	2255684.7 31	40265326. 9	3947044. 78
22 8	7	85.1288 4	840004.49	2254550.5 39	34606850. 9	4928665. 61
22 9	7	81.2554 4	678625.15	2526119.1 52	28065738. 2	4570385. 49
23 0	7	82.8116 1	902389.46	1904850.6 21	26998986. 7	4020959. 23
23 1	7	83.4063 4	156073.58	900091.1	22485755	2151751. 55
23 2	7	83.2535 4	914438.94	2201175.2 06	21430936. 5	59461.46 5
23 3	7	83.5398 9	961744.44	952767.904	3770364.1	2329108. 98
23 4	6	83.4747 2	772487.46	2981480.6 27	20293937. 7	3688927. 16

23 5	6	85.1578 3	 690412.99	634778.963	41159930. 3	585836.4 74
23 6	6	83.9336 3	664351.17	1486543.0 32	29264153. 4	1999926. 63
23 7	6	82.6949 7	774861.5	1217578.7 69	36349045. 6	3935233. 36
23 8	6	80.9521	103651.64	2547610.29	28140925. 5	779745.9 29
23 9	6	81.9396 3	945886.5	1630814.0 03	45479423. 3	5193703. 14
24 0	6	82.3811 1	462033.61	3059495.9 27	12110707. 6	1596343. 94
24 1	6	83.3475 1	580796.31	2710078.3 41	4594862.1	2000744. 43

24 2	6	82.9839	1049143.41	1509124.8 89	30177815. 6	2706812. 65
24 3	6	82.0859 8	986721.23	786095.575	25123869. 7	5428064. 72
24 4	6	80.7345 3	230009.63	1365030.6 46	17762089. 8	3989018. 54
24 5	6	82.2052 6	270120.12	505754.133	9584763.4	2222875. 94
24 6	6	82.1973 6	118775.11	4170767.6 98	47295224. 9	69577.39 8
24 7	6	82.6626 5	540183.82	3083244.2 12	23537307. 2	4108785. 23
24 8	6	83.1461 6	702758.32	3299655.5 46	1170131.4	4080377. 96
24 9	6	82.0508 7	213560.06	483451.385	27506938. 3	1148860. 42
25 0	6	80.7735 9	73144.15	108110.701	38107615. 7	1893315. 68
25 1	6	81.9500 3	733804.02	2860682.8 18	51685171. 3	4483835. 12
25 2	6	83.7724 3	437380.41	3948529.2 66	35036481. 8	1623587. 18
25 3	7	83.7098 2	1025260.29	3375340.7 51	4267370.4	3377511. 92
25 4	7	82.6394 5	156908.2	1352819.9 58	30926163. 8	9446.897
25 5	7	82.799 28	568426.85	3496451.7 99	17315516 .4	2850136.7 4
25 6	7	84.922 1	284365.61	119964.927	28926905 .7	3464107.4 4
25 7	7	83.443 61	92025.56	691462.404	38868167 .1	3903106.2 6
25 8	7	82.628 16	577486.2	2226773.1 07	6838763. 6	4680075.7 9

25 9	7	83.1194 5	238161.95	3179735.6 32	7871234. 6	4391949. 3
26 0	7	81.4922 6	703713.94	1703513.7 67	10197228. 2	2753317. 88
26 1	7	83.5719	964185.25	173117.275	20671767. 7	3807263. 09
26 2	7	85.7968	745040.93	761024.059	38582157. 2	2097520. 38
26 3	7	83.8031 7	443689.53	3584855.3 98	22158169. 4	2684134. 75
26 4	7	82.9318	855807.12	3579748.6 18	51757115. 7	2868105. 54
26 5	7	82.6034 7	1137262.22	2933227.4 65	14033400. 2	4185749. 72
26 6	7	80.6826 5	341899.69	4022778.0 52	34404797. 1	697822.2 98

26 7	783.3945 9	1169754.76	4212125.0 03	34238714. 3	2827136. 5
26 8	783.1541 3	1082384.08	1056647.8 27	14838174. 4	1106071. 61
26 9	782.6504 7	481415.96	3486467.8 82	2101134. 9	2663515. 07
27 0	781.3585 3	913673.6	61811.988	44428938. 1	4065947. 85
27 1	781.9892 8	1203704.01	3105694.0 46	15628244. 1	4182229. 66
27 2	782.8806 8	302825.12	3644499.5 99	33199078. 3	541033.6 11
27 3	786.1263	1134127.49	3997632.3 93	33979416. 7	2067895. 06

27 4	980.7715 5	332821.3	3427811.1 28	16371296. 2	5052526. 93
27 5	980.8888 5	62442.69	1982780.8 24	9407327. 8	5340105. 97
27 6	983.7944 9	925738.49	693501.556	9610551. 8	3828292. 88
27 7	981.0216 5	388732.63	368678.755	30004130. 1	3632216. 01
27 8	984.0644 2	370392.39	3222356.0 68	24015155. 2	2938275. 35
27 9	982.1673	60050.24	2233200.2 59	12855566. 4	3508848
28 0	983.6479 1	50902.58	2588040.3 99	7265053. 7	4546933. 71
28 1	984.3864 5	395405.46	462025.047	24387809. 6	4592257
28 2	983.7828 6	642798.47	1680267.4 61	16255169. 6	4584855. 99
28 3	984.2868 6	973107.52	2230074.1 33	34853119. 3	5459637. 64
28 4	983.2182 9	359358.36	1960264.3 07	34902857. 4	739480.9 64
28 5	985.3206 2	622980.97	1808297.52	25564092. 5	3604868. 08
28 6	982.4892 5	107462.22	1361323.1 36	3480338. 6	3535390. 52
28 7	982.7996 2	236161.33	1993227.3 24	13071725. 2	1417624. 19
28 8	983.9905 1	568465.45	4253858.5 15	3004631. 4	3419813. 85
28 9	983.3023 3	879398.98	3892819.1 31	45604350. 9	2076367. 02
29 0	983.7857 5	976108.51	3086380.7 24	45976917	2203147. 08
29 1	983.3855 9	720814.44	2676457.97	4734366	3983036. 09
29 2	983.3448 8	680451.54	357113.537	34467885. 4	1295516. 85

293	7	82.70173	408872.71	451000.472	47074090.6	2660984.45
294	7	84.04845	565370.41	1376975.921	676853	5358451.63
295	7	81.56268	537444.5	3347157.996	14595461.5	3997065.12
296	7	83.74976	952701.56	2928378.713	16890826.9	4167938.23
297	7	82.98392	1116979.42	918113.738	11019816.6	2278026.75
298	7	82.04645	366295.71	3307574.424	50570556.9	1309836.72
299	7	82.5939	596098.94	2768207.622	50449984.7	3800476.25
300	7	83.12855	654830.08	949033.596	1487381.4	1446292.86
303	7	81.33659	626887.19	1535792.94	51627144.6	4475277.91
301	7	84.57414	667355.8	4161709.929	43293061.6	2603546.03
304	7	82.64096	39874.23	153840.895	23547727.4	3511300.24
302	7	80.79402	191073.67	862205.143	40896636.1	887895.423
305	7	82.86258	299551.58	47965.456	14759655.8	5514417.95
306	7	83.40433	329617.39	961080.231	28465518.9	675634.81
307	7	83.56454	860246.09	1362560.353	38023045	4405423.36
308	7	81.91672	1021111.69	1614071.552	47464989.5	2200039.33
309	7	83.00303	271677.77	2380247.211	22854036.8	2830575.23
310	7	88.18147	57529.218	6814260.9	1162750	3087755
311	7	88.82126	113881.995	8338042.7	71804619	4734982.2
312	7	88.10815	36197.733	4049351.7	27281780	2496369.3
313	7	85.74068	14956.951	4903338.3	112920284	1353712.2
314	7	86.88837	131985.658	3840085.7	54810906	3809936.9
315	7	90.90149	143581.407	7526950.2	65157564	9520888.6
316	7	87.20899	30475.229	7106504.4	62740686	7091437.7
317	7	86.20344	118479.591	7613872.8	72482768	6877266
318	7	87.83032	110272.7	5675110.5	59200746	3210662.3
319	7	87.02104	107721.596	2836378.7	111281717	10327105.2
320	7	88.47132	74717.345	4004816.9	29077697	5225677.7
321	7	88.67773	15785.178	2688550.6	126579968	4989785
322	7	87.62832	92125.424	8191706.6	75842240	212354.6
323	7	90.77804	43607.025	3433756	1395954	9383740.1
324	7	89.14697	72408.06	8932922.2	134817952	4261790.9
325	7	88.95022	29387.476	1588144.9	28418913	10180621.8
326	7	88.35316	61663.767	8127159.5	49711145	9944369.7
327	7	87.65222	29442.597	304675.8	40708006	7568533
328	7	88.01796	83021.647	3247795	37382281	4644610.1
329	7	88.52633	4785.769	7345243.7	118206930	8640336.7
330	7	87.91572	39755.848	4256909.2	120574855	4742856.8

KNUST



331	7	88.08205	87622.152	8705914.3	127056705	7978300.4
332	7	87.00328	26540.937	1764029.2	103102523	10705478.1
333	7	88.21321	110152.115	2471658.3	32613252	6801890.3
334	7	88.34005	80230.831	6487277.6	103856678	9176406.1
335	7	87.99089	32266.654	7878471.9	64888924	6484969.8
336	7	88.72984	43003.87	3644291.1	16884711	1981714.3
337	7	89.34905	90827.068	2653690.3	65273776	11494811.1
338	7	89.2151	62324.722	5557371	56481462	4545928.1
339	7	88.36396	107390.436	2545526.7	47762861	11177873.8
340	7	86.08986	139280.753	473953.4	121456915	7791173
341	7	88.07091	124803.696	5717592.1	33522947	7551053.4
342	7	89.03802	15173.06	5703553.2	72641946	5685400.8
343	7	89.55549	107843.116	8209848.2	118997440	5290858.3
347	7	86.91763	48969.767	5966379.8	44904134	2919564.8
344	7	87.81817	30192.939	9000459.6	45091528	3784712.3
348	7	87.87288	67577.916	6360816.4	112181724	5523929.5
345	7	87.03665	8617.745	3460863.6	93788019	3962122
349	7	88.99133	17420.574	4995224.2	62520698	6522045.5
346	7	90.22766	22037.28	7126431.5	131187063	5576918.3
350	7	88.47752	139362.052	3946491.2	37116207	5877867.4
351	7	89.60176	102961.61	4439344.5	98843902	7990385.6
352	7	88.18573	144212.282	3900710.2	100545363	6429800.4
353	7	89.18348	4344.305	881263	49073372	9479460.7
354	7	87.62891	71287.979	5682420	90869199	9970019
355	7	89.27501	108662.873	4921103.4	27701986	8191991.6
356	7	90.68646	91382.941	4794873.5	99246894	5794972.4
357	7	88.79263	60789.409	3846473.1	129517864	641414
358	7	88.60031	27117.877	5934454	23382186	8449265.3
359	7	87.08	20960.724	1336310.1	101511023	5168022.9
360	7	87.75207	12140.289	5808940.7	126524290	927586.6
361	7	87.04723	89216.777	5209111.8	22888632	2212172.6
362	7	90.44098	50038.79	5164824.8	69377163	5098233.3
363	7	88.93036	68395.145	702072.1	12770840	209690.1
364	7	88.78139	64418.333	8916933.5	88467601	3681876.3
365	7	87.73192	113377.864	7793217.5	101701011	379160.9
366	7	87.76707	116895.776	8558184	88357120	1418711.7
367	7	88.21759	135790.951	8107327.1	49850693	6690671.9
368	7	88.96932	7347.509	556407.5	102653072	8604361.3

KNUST



369	7	88.41555	131118.104	4664984.3	123497500	1541643.1
370	7	87.59914	42452.078	8957092.3	101121441	2110555.7
371	7	86.84321	67025.852	6491318.7	108708979	9416577
372	7	87.74896	120673.903	1481114.5	93407662	9787958.8
373	7	90.39395	31545.729	4941682.2	93122299	7329519.8
374	7	88.37643	47275.584	4165454.6	126187617	9513223.4
375	7	87.52278	117837.67	1205160	51294475	5947601.9
376	7	88.70288	38989.361	7712295.1	123219883	7605241.7
377	7	88.326	47791.667	6254703.4	69884519	10028211.8
378	7	86.08146	129881.46	7154018.6	100072008	2129479
379	7	88.35381	56330.589	3522259.8	114008754	5772244.7
380	7	87.51088	49594.869	3718126.2	14452497	1047865.9
381	7	85.03622	130064.439	6466278.4	73361918	8042111.5
382	7	87.97838	97521.192	8693325.6	36235450	10301702.8
383	7	88.65683	11684.255	634660.4	112024280	3000582.9
384	7	87.82917	95230.442	2398947	134213045	5369305.9
385	7	86.88558	8939.539	5426567.6	105651387	9483516
386	7	87.9709	102281.938	1838620.2	90277797	4283052.2
387	7	80.93405	18893.393	2017793.96	71599335.7	11139355.6
388	7	81.78761	52628.423	2589259.12	79446784.3	9041513.74
389	7	81.75597	5410.698	3881184.51	76928301.4	8977516.94
390	7	82.26971	12433.328	4975336.7	72787222.5	12305899.4
391	7	82.37087	44735.318	2464417.2	64021247.8	12964730.4
392	7	78.64169	69232.352	2938030.65	22139599.1	8715117.16
393	7	79.60257	17487.007	1679134.33	3967099.4	8992386.86
394	7	80.52557	30102.742	1142224.73	130504096.3	8453196.84
395	7	80.72794	39645.416	4266073.01	53950406.1	12070358
396	7	81.91952	81112.667	4040858.77	58030283.9	4250493.71
397	7	80.71498	64015.098	3309232.69	45071763.8	4198286.42
398	7	82.33603	64708.074	1754929.96	127206574	7079372.27
399	7	80.31396	89276.39	4715484.72	110934347.6	10507609.3
400	7	80.16639	79915.014	4924611.74	38622027.7	11124397.6
401	7	81.67642	10993.11	2329276.19	22352279	3863904.88
402	7	81.06484	78848.642	4869448.16	77895161	1594868.86
403	7	80.40484	92107.396	2668479.6	24306423.2	12264077.5
404	7	81.59677	80823.793	4780490.72	134720666.4	12666307.8
405	7	81.41284	26051.429	3753404.06	136836955.9	5959751.84
406	7	79.66288	67535.018	394801.16	119377742.4	1676853.97

KNUST



40 7	781.0639 1	49488.862	1385248.3 8	21105952.1	9538501. 23
40 8	779.7295 6	62937.242	87926.55	41915171.6	783654.1 56
40 9	779.3397 6	6761.809	3102212.2 3	945606.7	10625524
41 0	781.3837 1	98673.492	133188.77	62179660.8	9539.223
41 1	780.2668 7	37376.315	3921553.3 5	32706981.3	12523469
41 2	780.2127 3	5488.605	4668480.1 4	42464145.7	12249290 .9

41 3	778.5527 3	13789.468	2570646.3	68350940.2	10132872 .1
41 4	780.2402 9	83057.407	866771.2	7857589.3	7494094. 56
41 5	781.4242 6	87305.369	1721122.7 6	101281739 .5	12642808 .1
41 6	780.3189 1	72778.83	1481315.0 9	6562230.9	1901217. 5
41 7	779.8220 1	62103.919	3312464.9 2	78973988.7	12496889 .8
41 8	781.5668 5	18758.782	3978923.8 3	40258216.6	7559841. 75
41 9	780.4811 1	55378.982	3404011.7 9	91459107.6	768163.5 17
42 0	780.5803 1	15946.829	361978.7	141652406	12140861 .8
42 1	779.8863 5	97923.347	1745684.1 5	58413457.3	678752.8 3
42 2	781.2211 1	72046.592	1289675.3 7	94456716.9	8421268
42 3	781.7583	91547.673	3122329.2 8	143599531 .7	4498293. 44
42 4	781.3776	15246.799	3733998.7 2	67825670.4	6333307. 84
42 5	781.7376 3	75846.882	2369335.6 2	45824409.5	1240084. 25

42 6	782.2091 2	10171.313	3009962.8 2	59152270.7	2171312. 52
42 7	780.1119 7	35109.236	3828456.2 6	142937323 .5	2572535. 77
42 8	879.3795 9	40963.836	3113795.7 2	103480415 .2	2841664. 87
42 9	880.3381 4	9241.361	450033.74	18488148.8	744584.7 41
43 0	878.4732 6	95514.649	2026933.7 2	121083286 .2	4517378. 89
43 1	879.4072 8	40734.913	4146713.8 5	38687936	12635199 .6
43 2	879.4687 9	54716.294	4746306.0 1	131641542 .9	2837121. 74

43 3	881.6153 7	33352.141	3495016	50978622.6	10030664 .1
43 4	879.1198 1	100984.657	234139.77	32532750.8	1372458. 7
43 5	879.2242 4	19722.699	4074335.8 4	134677194 .8	3781065. 14
43 6	879.5138 5	84235.523	4592595.7 1	143252394 .3	3689017. 37
43 7	881.8725 3	79224.163	2727703.8 8	68764019.2	2922656. 77
43 8	881.4153 9	97877.402	97074.85	82780806.7	6435749. 73
43 9	880.7227 6	30039.757	752668.68	58804667.1	7007667. 45
44 0	879.8170 2	100250.677	3401630.1 3	3981263.3	6819100. 49
44 1	880.7546 2	103683.672	3540653	100904004 .8	13065497 .8
44 2	782.0160 5	85055.573	3897491.6 1	1279402.4	10949571 .3
44 3	780.2817 2	37589.204	4571202.8 7	134881414 .7	13335066 .6
44 4	780.0472 4	86669.745	1232400.6 9	142651983 .1	6459629. 17

DMU	CCR Efficiency			Class			
	Stage 1 Efficiency	Stage 2 Efficiency	Overall Efficiency	Case 1	Case 2	Case 3	Case 4
1	0.51041	0.037362	0.863484	Class 4	Class 2	Class 4	Class 2
2	0.129048	0.008119	0.905399	Class 4	Class 2	Class 4	Class 2
3	0.189628	0.022087	0.965382	Class 4	Class 2	Class 4	Class 2

4	1	0.002175	1	Class 3	Class 3	Class 2	Class 2
5	0.216073	0.003382	0.877624	Class 4	Class 2	Class 4	Class 2
6	0.483117	0.020072	0.951268	Class 4	Class 2	Class 4	Class 2
7	0.563366	0.004981	0.918409	Class 4	Class 2	Class 4	Class 2
8	0.085544	0.012226	0.909131	Class 4	Class 2	Class 4	Class 2
9	0.355439	0.006896	1	Class 2	Class 2	Class 2	Class 2
10	0.472407	0.009612	0.930906	Class 4	Class 2	Class 4	Class 2
11	0.311048	0.022303	0.855815	Class 4	Class 2	Class 4	Class 2
12	0.038721	0.01963	0.884938	Class 4	Class 2	Class 4	Class 2
13	0.444092	0.001807	0.923815	Class 4	Class 2	Class 4	Class 2
14	0.301292	0.003506	0.887972	Class 4	Class 2	Class 4	Class 2
15	0.275436	0.002221	0.71688	Class 4	Class 4	Class 4	Class 4
16	0.278064	0.065517	0.629334	Class 4	Class 4	Class 4	Class 4
17	0.566731	0.005019	1	Class 2	Class 2	Class 2	Class 2
18	0.830898	0.003071	1	Class 2	Class 3	Class 2	Class 2
19	0.048307	0.026739	0.68221	Class 4	Class 4	Class 4	Class 4
20	0.1343	0.004237	0.661872	Class 4	Class 4	Class 4	Class 4

KNUST



**Appendix B: The CCR DEA efficiency and the classification of the 444 Bank Branches
(DMUs)**

21	0.185101	0.004791	0.783606	Class 4	Class 4	Class 4	Class 4
22	0.214974	0.037837	0.76041	Class 4	Class 4	Class 4	Class 4
23	0.446158	0.00145	0.839717	Class 4	Class 2	Class 4	Class 2
24	0.336537	0.008111	0.786179	Class 4	Class 4	Class 4	Class 4
25	0.028006	0.023187	0.757515	Class 4	Class 4	Class 4	Class 4
26	0.01182	0.092298	0.773928	Class 4	Class 4	Class 4	Class 4
27	0.318369	0.002069	0.808117	Class 4	Class 2	Class 4	Class 2
28	0.523699	0.00126	1	Class 2	Class 2	Class 2	Class 2
29	0.314873	0.002396	0.813951	Class 4	Class 2	Class 4	Class 2
30	0.2881	0.002338	0.807853	Class 4	Class 2	Class 4	Class 2
31	0.469401	0.001427	0.931386	Class 4	Class 2	Class 4	Class 2
32	0.329428	0.002892	0.798275	Class 4	Class 4	Class 4	Class 4
33	0.556094	0.00148	1	Class 2	Class 2	Class 2	Class 2
34	0.367568	0.00181	0.827586	Class 4	Class 2	Class 4	Class 2
38	0.35393	0.018622	0.893469	Class 4	Class 2	Class 4	Class 2
35	0.233879	0.012344	0.87085	Class 4	Class 2	Class 4	Class 2
39	0.444926	0.012937	1	Class 2	Class 2	Class 2	Class 2
36	0.42567	0.039589	0.923673	Class 4	Class 2	Class 4	Class 2
40	0.204714	0.00441	0.888666	Class 4	Class 2	Class 4	Class 2
37	0.115662	0.039492	0.907363	Class 4	Class 2	Class 4	Class 2
41	0.274533	0.003844	0.914323	Class 4	Class 2	Class 4	Class 2
42	1	0.018177	1	Class 3	Class 3	Class 2	Class 2
43	0.431135	0.001747	0.919634	Class 4	Class 2	Class 4	Class 2
44	0.423596	0.043561	0.892051	Class 4	Class 2	Class 4	Class 2
45	0.610891	0.001262	1	Class 2	Class 2	Class 2	Class 2
46	0.288549	0.002862	0.96554	Class 4	Class 2	Class 4	Class 2
47	0.218608	0.00351	0.901991	Class 4	Class 2	Class 4	Class 2
48	0.394909	0.052855	0.863667	Class 4	Class 2	Class 4	Class 2
49	0.497868	0.001534	0.943356	Class 4	Class 2	Class 4	Class 2
50	0.426821	0.046258	0.86984	Class 4	Class 2	Class 4	Class 2
51	1	0.043299	1	Class 3	Class 3	Class 2	Class 2
52	0.555211	0.038674	0.864187	Class 4	Class 2	Class 4	Class 2
53	0.274186	0.067967	0.846584	Class 4	Class 2	Class 4	Class 2
54	0.346354	0.00223	0.931909	Class 4	Class 2	Class 4	Class 2
55	0.29399	0.066705	0.83778	Class 4	Class 2	Class 4	Class 2
56	0.071285	0.010776	0.871243	Class 4	Class 2	Class 4	Class 2
57	0.092978	0.007115	0.755028	Class 4	Class 4	Class 4	Class 4
58	0.415331	0.001605	0.818824	Class 4	Class 2	Class 4	Class 2
59	0.444651	0.001506	0.933609	Class 4	Class 2	Class 4	Class 2
60	0.235273	0.002911	0.830839	Class 4	Class 2	Class 4	Class 2

KNUST



61	0.390082	0.003295	0.922585	Class 4	Class 2	Class 4	Class 2
62	0.356193	0.001865	0.817932	Class 4	Class 2	Class 4	Class 2
63	0.336715	0.00197	0.818176	Class 4	Class 2	Class 4	Class 2
64	0.240565	0.00276	0.797182	Class 4	Class 4	Class 4	Class 4
65	0.021206	0.045055	0.835801	Class 4	Class 2	Class 4	Class 2
66	0.447906	0.001482	0.847057	Class 4	Class 2	Class 4	Class 2
67	0.12892	0.005202	0.80453	Class 4	Class 2	Class 4	Class 2
68	0.096667	0.006839	0.759633	Class 4	Class 4	Class 4	Class 4
69	0.271929	0.003249	0.83642	Class 4	Class 2	Class 4	Class 2
70	0.055687	0.012085	0.756909	Class 4	Class 4	Class 4	Class 4
71	0.610006	0.03609	0.874879	Class 4	Class 2	Class 4	Class 2
72	1	0.001433	1	Class 3	Class 3	Class 2	Class 2
73	0.05434	0.018299	0.828893	Class 4	Class 2	Class 4	Class 2
74	0.103459	0.005673	0.645367	Class 4	Class 4	Class 4	Class 4
75	0.141514	0.006259	1	Class 2	Class 2	Class 2	Class 2
76	0.061679	0.00941	0.65225	Class 4	Class 4	Class 4	Class 4
77	0.253551	0.070107	0.6186	Class 4	Class 4	Class 4	Class 4
78	0.192286	0.005107	0.646839	Class 4	Class 4	Class 4	Class 4
79	0.026666	0.000343	0.695559	Class 4	Class 4	Class 4	Class 4
80	0.043352	0.007293	0.746034	Class 4	Class 4	Class 4	Class 4
81	0.030126	0.053698	0.999251	Class 4	Class 2	Class 2	Class 2
86	0.297222	0.061168	0.630784	Class 4	Class 4	Class 4	Class 4
87	0.421406	0.037496	0.604399	Class 4	Class 4	Class 4	Class 4
88	0.406552	0.009652	1	Class 2	Class 2	Class 2	Class 2
89	0.127924	0.153432	0.708022	Class 4	Class 4	Class 4	Class 4
90	0.331269	0.002305	0.912381	Class 4	Class 2	Class 4	Class 2
91	0.849506	0.001279	1	Class 2	Class 3	Class 2	Class 2
92	0.032601	0.023342	0.853471	Class 4	Class 2	Class 4	Class 2
93	0.784393	0.023617	1	Class 2	Class 2	Class 2	Class 2
94	0.438281	0.041797	0.860842	Class 4	Class 2	Class 4	Class 2
95	0.270714	0.064075	0.848983	Class 4	Class 2	Class 4	Class 2
96	0.300407	0.063149	0.853652	Class 4	Class 2	Class 4	Class 2
97	0.499953	0.036739	0.861094	Class 4	Class 2	Class 4	Class 2
98	0.354291	0.054253	0.857441	Class 4	Class 2	Class 4	Class 2
99	0.154178	0.004989	0.898771	Class 4	Class 2	Class 4	Class 2
100	0.059928	0.012563	0.851934	Class 4	Class 2	Class 4	Class 2

KNUST



101	0.403142	0.044161	0.909206	Class 4	Class 2	Class 4	Class 2
102	0.040252	0.019337	0.871198	Class 4	Class 2	Class 4	Class 2
103	0.23053	0.078954	0.894393	Class 4	Class 2	Class 4	Class 2
104	0.10392	0.007379	0.865807	Class 4	Class 2	Class 4	Class 2
105	0.196752	0.003858	0.924014	Class 4	Class 2	Class 4	Class 2
106	0.332308	0.061316	0.866975	Class 4	Class 2	Class 4	Class 2
107	0.50235	0.040183	0.829004	Class 4	Class 2	Class 4	Class 2
108	0.566247	0.00139	0.96277	Class 4	Class 2	Class 4	Class 2
109	1	0.024499	1	Class 3	Class 3	Class 2	Class 2
110	0.366391	0.046532	0.930789	Class 4	Class 2	Class 4	Class 2
111	0.019503	0.065577	1	Class 2	Class 2	Class 2	Class 2
112	0.37688	0.047941	0.865	Class 4	Class 2	Class 4	Class 2
113	0.585254	0.03208	0.849674	Class 4	Class 2	Class 4	Class 2
114	0.422802	0.044828	0.877959	Class 4	Class 2	Class 4	Class 2
115	0.181763	0.004244	0.900455	Class 4	Class 2	Class 4	Class 2
116	0.198369	0.004369	0.933981	Class 4	Class 2	Class 4	Class 2
117	0.424176	0.04194	0.849459	Class 4	Class 2	Class 4	Class 2
118	0.212278	0.009815	0.923333	Class 4	Class 2	Class 4	Class 2
119	0.206058	0.003553	0.874061	Class 4	Class 2	Class 4	Class 2
120	0.441629	0.003201	0.872267	Class 4	Class 2	Class 4	Class 2
121	0.580752	0.00124	1	Class 2	Class 2	Class 2	Class 2
126	0.462025	0.003135	0.876338	Class 4	Class 2	Class 4	Class 2
122	0.058686	0.024294	0.873205	Class 4	Class 2	Class 4	Class 2
127	0.632826	0.002867	0.914927	Class 4	Class 2	Class 4	Class 2
123	0.601164	0.00344	0.884119	Class 4	Class 2	Class 4	Class 2
128	0.447556	0.004341	0.917541	Class 4	Class 2	Class 4	Class 2
124	0.837549	0.001169	1	Class 2	Class 3	Class 2	Class 2
129	0.767617	0.001589	1	Class 2	Class 2	Class 2	Class 2
125	0.041549	0.018476	0.856316	Class 4	Class 2	Class 4	Class 2
130	0.247678	0.004582	0.988177	Class 4	Class 2	Class 4	Class 2
131	0.314942	0.002769	0.988611	Class 4	Class 2	Class 4	Class 2
132	0.904469	0.000928	1	Class 2	Class 3	Class 2	Class 2
133	0.396777	0.002965	0.989482	Class 4	Class 2	Class 4	Class 2
134	0.502856	0.001742	1	Class 2	Class 2	Class 2	Class 2
135	0.615602	0.002443	0.992526	Class 4	Class 2	Class 4	Class 2
136	0.677172	0.003027	1	Class 2	Class 2	Class 2	Class 2
137	0.284082	0.006567	1	Class 2	Class 2	Class 2	Class 2
138	0.052723	0.034683	1	Class 2	Class 2	Class 2	Class 2
139	0.377066	0.005122	0.993743	Class 4	Class 2	Class 4	Class 2
140	0.620313	0.002433	1	Class 2	Class 2	Class 2	Class 2

KNUST




141	0.555336	0.001538	1	Class 2	Class 2	Class 2	Class 2
142	0.894108	0.002015	1	Class 2	Class 3	Class 2	Class 2
143	0.278496	0.005875	1	Class 2	Class 2	Class 2	Class 2
144	0.634499	0.001395	1	Class 2	Class 2	Class 2	Class 2
145	0.694601	0.003554	1	Class 2	Class 2	Class 2	Class 2
146	0.806759	0.00103	1	Class 2	Class 3	Class 2	Class 2
147	0.750817	0.00227	1	Class 2	Class 2	Class 2	Class 2
148	1	0.002249	1	Class 3	Class 3	Class 2	Class 2
149	0.249434	0.003395	1	Class 2	Class 2	Class 2	Class 2
150	1	0.002471	1	Class 3	Class 3	Class 2	Class 2
151	0.44025	0.002045	1	Class 2	Class 2	Class 2	Class 2
152	0.142928	0.007969	0.822753	Class 4	Class 2	Class 4	Class 2
153	0.362185	0.003175	0.871956	Class 4	Class 2	Class 4	Class 2
154	0.002284	0.450441	0.87093	Class 4	Class 2	Class 4	Class 2
155	0.07089	0.011561	0.852999	Class 4	Class 2	Class 4	Class 2
156	0.301519	0.002462	0.858944	Class 4	Class 2	Class 4	Class 2
157	0.234344	0.003197	0.861134	Class 4	Class 2	Class 4	Class 2
158	0.275184	0.002632	0.84212	Class 4	Class 2	Class 4	Class 2
159	0.061079	0.012061	0.847198	Class 4	Class 2	Class 4	Class 2
160	0.234374	0.003109	0.836424	Class 4	Class 2	Class 4	Class 2
161	0.225179	0.003507	0.828976	Class 4	Class 2	Class 4	Class 2
162	0.120276	0.013449	0.822938	Class 4	Class 2	Class 4	Class 2
163	0.202576	0.003775	0.857827	Class 4	Class 2	Class 4	Class 2
164	0.002465	0.307508	0.825274	Class 4	Class 2	Class 4	Class 2
165	0.184404	0.004368	0.858685	Class 4	Class 2	Class 4	Class 2
166	0.363466	0.002544	0.853796	Class 4	Class 2	Class 4	Class 2
167	0.670048	0.003664	0.852658	Class 4	Class 2	Class 4	Class 2
168	0.296492	0.002418	0.869911	Class 4	Class 2	Class 4	Class 2
169	0.023335	0.004279	0.738749	Class 4	Class 4	Class 2	Class 2
175	0.015304	0.055576	0.73693	Class 4	Class 4	Class 4	Class 4
176	0.352062	0.006137	0.72641	Class 4	Class 4	Class 4	Class 4
177	0.224608	0.003486	0.729681	Class 4	Class 4	Class 4	Class 4
178	0.316832	0.002153	0.833344	Class 4	Class 2	Class 4	Class 2
179	0.186464	0.003492	0.764345	Class 4	Class 4	Class 4	Class 4
180	0.113557	0.005693	0.724325	Class 4	Class 4	Class 4	Class 4

KNUST



18 1	0.49629 9	0.00223 1	0.88232 3	Class 4	Clas s 2	Clas s 4	Clas s 2
18 2	0.03649 5	0.02247 3	0.73269 2	Class 4	Clas s 4	Clas s 4	Clas s 4
18 3	0.19716 1	0.00469 9	0.75253 5	Class 4	Clas s 4	Clas s 4	Clas s 4
18 4	0.30160 9	0.03170 8	0.71908 6	Class 4	Clas s 4	Clas s 4	Clas s 4
18 5	0.19793 5	0.00510 3	0.72728 4	Class 4	Clas s 4	Clas s 4	Clas s 4


18 6	0.26565 2	0.00241 1	0.75718 4	Class 4	Clas s 4	Clas s 4	Clas s 4
18 7	0.47664 4	0.03146 7	0.75708 6	Class 4	Clas s 4	Clas s 4	Clas s 4
18 8	0.28933 9	0.00350 8	0.74351 4	Class 4	Clas s 4	Clas s 4	Clas s 4
18 9	0.27473 1	0.00266 8	0.81596 8	Class 4	Clas s 2	Clas s 4	Clas s 2
19 0	0.10777 4	0.00577 5	0.71677 3	Class 4	Clas s 4	Clas s 4	Clas s 4
19 1	0.27817 5	0.03780 2	0.73940 6	Class 4	Clas s 4	Clas s 4	Clas s 4
19 2	0.12608 9	0.00527 2	0.65648 7	Class 4	Clas s 4	Clas s 4	Clas s 4
19 3	0.24201 5	0.00278 7	0.65362 4	Class 4	Clas s 4	Clas s 4	Clas s 4
19 4	0.15729 7	0.00367 5	0.66997 2	Class 4	Clas s 4	Clas s 4	Clas s 4
19 5	0.12587 7	0.00457 6	0.65434 5	Class 4	Clas s 4	Clas s 4	Clas s 4
19 6	0.01898 7	0.03156 5	0.65291 8	Class 4	Clas s 4	Clas s 4	Clas s 4
19 7	0.28615 7	0.00202 3	0.74554 9	Class 4	Clas s 4	Clas s 4	Clas s 4
19 8	0.20051	0.02238 2	0.65041 4	Class 4	Clas s 4	Clas s 4	Clas s 4
19 9	0.11767 6	0.00482 3	0.64572	Class 4	Clas s 4	Clas s 4	Clas s 4
20 0	0.00057 2	1	0.65058 5	Class 4	Clas s 4	Clas s 1	Clas s 1

201	0.191786	0.00306	0.66515	 <p>Class 4</p>	Class 4	Class 4	Class 4
202	0.112304	0.004999	0.641051	Class 4	Class 4	Class 4	Class 4
203	0.243818	0.002957	0.838088	Class 4	Class 2	Class 4	Class 2
204	0.360095	0.002697	0.841563	Class 4	Class 2	Class 4	Class 2
205	0.236013	0.003162	0.82673	Class 4	Class 2	Class 4	Class 2
206	0.205833	0.00348	0.823992	Class 4	Class 2	Class 4	Class 2

20 7	0.02615 2	0.02799 1	0.82496 7	Class 4	Class s 2	Class s 4	Class s 2
20 8	0.07358	0.01024	0.82654 2	Class 4	Class s 2	Class s 4	Class s 2
20 9	0.07120 5	0.01040 5	0.83819 5	Class 4	Class s 2	Class s 4	Class s 2
21 0	0.18537 3	0.00465 7	0.90712 6	Class 4	Class s 2	Class s 4	Class s 2
21 1	0.24775 2	0.00427 4	0.83478	Class 4	Class s 2	Class s 4	Class s 2
21 2	0.24270 1	0.00295 8	0.84013 4	Class 4	Class s 2	Class s 4	Class s 2
21 3	0.27775 6	0.00264	0.85426 5	Class 4	Class s 2	Class s 4	Class s 2

21 4	0.19209 3	0.01018 6	0.90989 6	Class 4	Class 2	Class 4	Class 2
21 5	0.06476 4	0.01806 1	0.81981 8	Class 4	Class 2	Class 4	Class 2
21 6	0.07190 8	0.01390 6	0.83309	Class 4	Class 2	Class 4	Class 2
21 7	0.31236 2	0.00233 4	0.83786 9	Class 4	Class 2	Class 4	Class 2
21 8	0.19245 5	0.00382 6	0.84446 4	Class 4	Class 2	Class 4	Class 2
21 9	0.00120 5	0.94456 5	0.85730 3	Class 4	Class 2	Class 4	Class 3
22 0	0.17794 1	0.0044	0.87486	Class 4	Class 2	Class 4	Class 2
22 1	0.06813 7	0.01079 3	0.84015	Class 4	Class s 2	Class s 4	Class s 2
22 2	0.20982 9	0.00351 8	0.84090 4	Class 4	Class s 2	Class s 4	Class s 2
22 3	0.41748 9	0.00487 6	0.83268 7	Class 4	Class s 2	Class s 4	Class s 2
22 4	0.18913 6	0.00483 7	0.85213 9	Class 4	Class s 2	Class s 4	Class s 2
22 5	0.25868 3	0.00289 7	0.86770 8	Class 4	Class s 2	Class s 4	Class s 2
22 6	0.15393 8	0.03642 1	1	Class 2	Class s 2	Class s 2	Class s 2
22 7	0.27648 8	0.00320 5	0.82129 5	Class 4	Class s 2	Class s 4	Class s 2
22 8	0.23388 2	0.00465 6	0.85524 2	Class 4	Class s 2	Class s 4	Class s 2
22 9	0.19083 7	0.00532 4	0.81712 5	Class 4	Class s 2	Class s 4	Class s 2

23 0	0.18218 3	0.00486 9	0.83150 5	Class 4	Clas s 2	Clas s 4	Clas s 2
23 1	0.15737 1	0.01199 5	0.85673 3	Class 4	Clas s 2	Clas s 4	Clas s 2
23 2	0.14439 6	0.00498 9	0.83544 6	Class 4	Clas s 2	Clas s 4	Clas s 2
23 3	0.02559 9	0.02845 6	0.83861 7	Class 4	Clas s 2	Clas s 4	Clas s 2
23 4	0.16101 7	0.00594 3	0.97629 6	Class 4	Clas s 2	Clas s 4	Clas s 2
23 5	0.33057 3	0.04813	0.99575 5	Class 4	Clas s 2	Clas s 4	Clas s 2
23 6	0.23178 3	0.00903	0.99054 7	Class 4	Clas s 2	Clas s 4	Clas s 2
23 7	0.28658 9	0.00354	0.97675	Class 4	Clas s 2	Clas s 4	Clas s 2
23 8	0.22844 6	0.00369 4	0.98181 2	Class 4	Clas s 2	Clas s 4	Clas s 2
23 9	0.35495 3	0.00373 4	0.95919 6	Class 4	Clas s 2	Clas s 4	Clas s 2
24 0	0.09716 4	0.00873 6	0.96496 9	Class 4	Clas s 2	Clas s 4	Clas s 2


24 1	0.03655 6	0.02329 6	0.97587 3	 <p>Class 4</p>	Class s 2	Class s 4	Class s 2
24 2	0.23442 8	0.00519	0.97765 9	Class 4	Class s 2	Class s 4	Class s 2
24 3	0.19946 9	0.00706 4	0.96493 8	Class 4	Class s 2	Class s 4	Class s 2
24 4	0.14391 7	0.00734 2	0.94925 2	Class 4	Class s 2	Class s 4	Class s 2
24 5	0.07781 6	0.02237 6	0.96551 6	Class 4	Class s 2	Class s 4	Class s 2
24 6	0.38945 7	0.00223 2	1	Class 2	Class s 2	Class s 2	Class s 2

24 7	0.18841 6	0.00570 7	0.96937 8	Class 4	Clas s 2	Clas s 4	Clas s 2
24 8	0.00934 3	0.11400 8	0.97362 9	Class 4	Clas s 2	Clas s 4	Clas s 2
24 9	0.22400 2	0.00383 1	1	Class 2	Clas s 2	Clas s 2	Clas s 2
25 0	0.50801	0.00575 6	1	Class 2	Clas s 2	Clas s 2	Clas s 2
25 1	0.40996 1	0.00283 6	0.96005 5	Class 4	Clas s 2	Clas s 4	Clas s 2
25 2	0.28461 8	0.01619 3	0.98052 1	Class 4	Clas s 2	Clas s 4	Clas s 2
25 3	0.02888 9	0.02587 6	0.83966 9	Class 4	Clas s 2	Clas s 4	Clas s 2
25 4	0.21602 6	0.00343 2	0.87227 9	Class 4	Clas s 2	Clas s 4	Clas s 2
25 5	0.11888 2	0.00614 1	0.83109 6	Class 4	Clas s 2	Clas s 4	Clas s 2
25 6	0.30442 7	0.00391 5	0.91767 6	Class 4	Clas s 2	Clas s 4	Clas s 2
25 7	0.27305 7	0.00328 3	0.90400 2	Class 4	Clas s 2	Clas s 4	Clas s 2

258	0.046753	0.022374	0.83153	Class 4	Class 2	Class 4	Class 2
259	0.054518	0.018242	0.850865	Class 4	Class 2	Class 4	Class 2
260	0.06945	0.010263	0.81828	Class 4	Class 2	Class 4	Class 2

26 1	0.17330 3	0.00602 1	0.84567 4	Class 4	Clas s 2	Clas s 4	Clas s 2
26 2	0.26524 6	0.00285 6	0.86723 7	Class 4	Clas s 2	Clas s 4	Clas s 2
26 3	0.15281 4	0.00485 7	0.84256 4	Class 4	Clas s 2	Clas s 4	Clas s 2
26 4	0.35282 9	0.00205 8	0.83624 6	Class 4	Clas s 2	Clas s 4	Clas s 2
26 5	0.09425 4	0.00975 2	0.82749 8	Class 4	Clas s 2	Clas s 4	Clas s 2
26 6	0.23912 9	0.00301 2	0.81766 1	Class 4	Clas s 2	Clas s 4	Clas s 2
26 7	0.23302 2	0.00312 8	0.83953 6	Class 4	Clas s 2	Clas s 4	Clas s 2
26 8	0.10007 1	0.00719 7	0.89328 8	Class 4	Clas s 2	Clas s 4	Clas s 2
26 9	0.01445 9	0.05051 8	0.82977 5	Class 4	Clas s 2	Clas s 4	Clas s 2


27 0	0.57219	0.00299 2	1	Class 2	Clas s 2	Clas s 2	Clas s 2
27 1	0.10497 1	0.00874 9	0.82184 2	Class 4	Clas s 2	Clas s 4	Clas s 2
27 2	0.23002 4	0.00320 6	0.84321 4	Class 4	Clas s 2	Clas s 4	Clas s 2
27 3	0.23091 5	0.00325 5	0.87018 6	Class 4	Clas s 2	Clas s 4	Clas s 2
27 4	0.08815	0.01009	0.64363 4	Class 4	Clas s 4	Clas s 4	Clas s 4
27 5	0.05135 1	0.01855 9	0.67228	Class 4	Clas s 4	Clas s 4	Clas s 4
27 6	0.05214 6	0.01302 3	0.65866 7	Class 4	Clas s 4	Clas s 4	Clas s 4
27 7	0.17580 5	0.00395 8	0.65202 3	Class 4	Clas s 4	Clas s 4	Clas s 4
27 8	0.12921 1	0.00776 3	0.65849 8	Class 4	Clas s 4	Clas s 4	Clas s 4
27 9	0.07015 9	0.00892 4	0.67402 8	Class 4	Clas s 4	Clas s 4	Clas s 4
28 0	0.03965 7	0.02046 2	0.68888 1	Class 4	Clas s 4	Clas s 4	Clas s 4

28 1	0.13420 2	0.00615 6	0.67337 9	 <p>Class 4</p>	Class 4	Class 4	Class 4
28 2	0.08699 5	0.00922 2	0.65603 3	Class 4	Class 4	Class 4	Class 4
28 3	0.18417 8	0.00518 6	0.65959 8	Class 4	Class 4	Class 4	Class 4
28 4	0.18848 3	0.00306 2	0.67591 1	Class 4	Class 4	Class 4	Class 4
28 5	0.13685 9	0.00461	0.66960 6	Class 4	Class 4	Class 4	Class 4
28 6	0.01899 3	0.03321 1	0.66993 2	Class 4	Class 4	Class 4	Class 4

28 7	0.07088 6	0.00813 5	0.65205 6	Class 4	Class s 4	Class s 4	Class s 4
28 8	0.01606 8	0.03723 6	0.65675 4	Class 4	Class s 4	Class s 4	Class s 4
28 9	0.24148 2	0.00815 9	0.64980 1	Class 4	Class s 4	Class s 4	Class s 4
29 0	0.24240 5	0.00762 2	0.65399 6	Class 4	Class s 4	Class s 4	Class s 4
29 1	0.02520 5	0.02764 7	0.65199	Class 4	Class s 4	Class s 4	Class s 4
29 2	0.20359 4	0.00310 5	0.66870 8	Class 4	Class s 4	Class s 4	Class s 4
29 3	0.33034 1	0.0038	0.83590 3	Class 4	Class s 2	Class s 4	Class s 2
29 4	0.00464 3	0.25882 9	0.84662 5	Class 4	Class s 2	Class s 4	Class s 2
29 5	0.10012 8	0.00895 3	0.82030 6	Class 4	Class s 2	Class s 4	Class s 2
29 6	0.11401 7	0.00806 7	0.83982 6	Class 4	Class s 2	Class s 4	Class s 2
29 7	0.07498 7	0.00967 1	0.85338 4	Class 4	Class s 2	Class s 4	Class s 2
29 8	0.34870 7	0.00208 4	0.83820 9	Class 4	Class s 2	Class s 4	Class s 2
29 9	0.34423	0.00246 3	0.83208 7	Class 4	Class s 2	Class s 4	Class s 2
30 0	0.01018 1	0.07177 7	0.83308 1	Class 4	Class s 2	Class s 4	Class s 2
30 1	0.29861 6	0.00250 9	0.84983 3	Class 4	Class 2	Class 4	Class 2

30 2	0.28582 5	0.04214 9	0.81089 8	Class 4	Class s 2	Class s 4	Class s 2
30 3	0.35299 3	0.00283 4	0.82032 7	Class 4	Class s 2	Class s 4	Class s 2
30 4	0.26216 9	0.00487 5	0.94981 3	Class 4	Class s 2	Class s 4	Class s 2
30 5	0.21109 2	0.01221 5	1	Class 2	Class s 2	Class s 2	Class s 2
30 6	0.19764	0.00376 3	0.84647 5	Class 4	Class s 2	Class s 4	Class s 2
30 7	0.25750 8	0.00378 8	0.84619 5	Class 4	Class s 2	Class s 4	Class s 2
30 8	0.31895 2	0.00221 6	0.82718	Class 4	Class s 2	Class s 4	Class s 2
30 9	0.15827 8	0.00466 4	0.84020 5	Class 4	Class s 2	Class s 4	Class s 2


31 0	0.00841 4	0.12663 2	0.91113 6	Class 4	Clas s 2	Clas s 4	Clas s 2
31 1	0.49957	0.00215 6	0.95373 4	Class 4	Clas s 2	Clas s 4	Clas s 2
31 2	0.19145 9	0.00905 2	0.91573	Class 4	Clas s 2	Clas s 4	Clas s 2
31 3	0.86233 7	0.00936 9	0.94031 4	Class 4	Clas s 3	Clas s 4	Clas s 2
31 4	0.38245 3	0.00227 3	0.92823 3	Class 4	Clas s 2	Clas s 4	Clas s 2
31 5	0.47426 3	0.00477 7	0.98678 1	Class 4	Clas s 2	Clas s 4	Clas s 2
31 6	0.43989 5	0.00369 5	0.96056 5	Class 4	Clas s 2	Clas s 4	Clas s 2
31 7	0.52851 7	0.00310 2	0.92917 5	Class 4	Clas s 2	Clas s 4	Clas s 2
31 8	0.42210 3	0.00227 1	0.93708 6	Class 4	Clas s 2	Clas s 4	Clas s 2
31 9	0.77513 5	0.00303 4	0.98684 8	Class 4	Clas s 2	Clas s 4	Clas s 2

320	0.203606	0.005876	0.941184	 <p>Class 4</p>	Class 2	Class 4	Class 2
321	0.927882	0.001289	1	Class 2	Class 3	Class 2	Class 2
322	0.528138	0.001484	0.953898	Class 4	Class 2	Class 4	Class 2
323	0.009765	0.219772	1	Class 2	Class 2	Class 2	Class 2
324	0.940317	0.001034	1	Class 2	Class 3	Class 2	Class 2
325	0.199839	0.011712	1	Class 2	Class 2	Class 2	Class 2

32 6	0.34702 6	0.00654	0.98188 1	Class 4	Class s 2	Class s 4	Class s 2
32 7	0.31717 1	0.00607 9	1	Class 2	Class s 2	Class s 2	Class s 2
32 8	0.26052 2	0.00406 2	0.93836	Class 4	Class s 2	Class s 4	Class s 2
32 9	1	0.00239	1	Class 3	Class s 3	Class s 2	Class s 2
33 0	0.84755 5	0.00128 6	0.98353	Class 4	Class s 3	Class s 4	Class s 2
33 1	0.88508 3	0.00205 3	0.98704 5	Class 4	Class s 3	Class s 4	Class s 2
33 2	0.72497 3	0.00339 5	1	Class 2	Class s 2	Class s 2	Class s 2
33 3	0.22742 4	0.00681 9	0.94042	Class 4	Class s 2	Class s 4	Class s 2
33 4	0.74817 7	0.00288 9	0.99228 6	Class 4	Class s 2	Class s 4	Class s 2
33 5	0.45457 4	0.00326 7	0.96595 4	Class 4	Class s 2	Class s 4	Class s 2
33 6	0.11807	0.02005 2	0.91434 8	Class 4	Class s 2	Class s 4	Class s 2
33 7	0.45528 9	0.00575 7	1	Class 2	Class s 2	Class s 2	Class s 2
33 8	0.40272 9	0.00263 1	0.96506	Class 4	Class s 2	Class s 4	Class s 2
33 9	0.33302 5	0.00765 1	0.98404 4	Class 4	Class s 2	Class s 4	Class s 2
34 0	0.85228 1	0.00209 7	1	Class 2	Class s 3	Class s 2	Class s 2
34 1	0.23903 6	0.00736 4	0.95098	Class 4	Class 2	Class 4	Class 2
34 2	0.56162 1	0.00255 9	0.98982 7	Class 4	Class 2	Class 4	Class 2
34 3	0.82811 5	0.00145 4	0.97592 9	Class 4	Class 3	Class 4	Class 2
34 4	0.31619 3	0.00274 4	0.95800 8	Class 4	Class 2	Class 4	Class 2
34 5	0.77991 2	0.00138 1	1	Class 2	Class 2	Class 2	Class 2


34 6	0.92345 7	0.00139	1	Class 2	Class s 3	Class s 2	Class s 2
34 7	0.32252 1	0.00364 5	0.93568 4	Class 4	Class s 2	Class s 4	Class s 2
34 8	0.80725 9	0.00161	0.96238 3	Class 4	Class s 3	Class s 4	Class s 2

34 9	0.47174 9	0.00341 1	0.98219 4	Class 4	Clas s 2	Clas s 4	Clas s 2
35 0	0.25924 2	0.00517 8	0.93430 5	Class 4	Clas s 2	Clas s 4	Clas s 2
35 1	0.69498 5	0.00264 3	0.97906 9	Class 4	Clas s 2	Clas s 4	Clas s 2
35 2	0.70181	0.00209 1	0.95431	Class 4	Clas s 2	Clas s 4	Clas s 2
35 3	0.62419	0.00631 5	1	Class 2	Clas s 2	Clas s 2	Clas s 2
35 4	0.64869 2	0.00358 7	0.99201 2	Class 4	Clas s 2	Clas s 4	Clas s 2
35 5	0.19581 9	0.00966 8	0.97004 5	Class 4	Clas s 2	Clas s 4	Clas s 2
35 6	0.70089 1	0.00190 9	0.98757	Class 4	Clas s 2	Clas s 4	Clas s 2
35 7	0.90563 7	0.01865 6	0.91120 9	Class 4	Clas s 3	Clas s 4	Clas s 2
35 8	0.17256 7	0.01181 4	0.97779 1	Class 4	Clas s 2	Clas s 4	Clas s 2
35 9	0.71557 5	0.00166 4	0.98695 7	Class 4	Clas s 2	Clas s 4	Clas s 2

36 0	0.99692 7	0.01308 7	0.96464 1	 Class 4	Class 3	Class 4	Class 2
36 1	0.16242 2	0.01237 1	0.91187 1	Class 4	Class 2	Class 4	Class 2
36 2	0.49262 1	0.00240 3	0.98627 8	Class 4	Class 2	Class 4	Class 2
36 3	0.08980 8	0.00894 3	0.93734 2	Class 4	Class 2	Class 4	Class 2
36 4	0.61744	0.00136 1	0.97682 6	Class 4	Class 2	Class 4	Class 2
36 5	0.70758 5	0.00110 8	1	Class 2	Class 2	Class 2	Class 2

36 6	0.61465 5	0.00127 6	0.98143 6	Class 4	Class s 2	Class s 4	Class s 2
36 7	0.34651 4	0.00438 8	0.93916 5	Class 4	Class s 2	Class s 4	Class s 2
36 8	1	0.00274	1	Class 3	Class s 3	Class s 2	Class s 2
36 9	0.86986 1	0.00796 9	0.91081 4	Class 4	Class s 3	Class s 4	Class s 2
37 0	0.70702 7	0.00464 4	0.93458 6	Class 4	Class s 2	Class s 4	Class s 2
37 1	0.78347 4	0.00283 2	0.97584 2	Class 4	Class s 2	Class s 4	Class s 2
37 2	0.65324 6	0.00342 6	0.97246 2	Class 4	Class s 2	Class s 4	Class s 2
37 3	0.66936 9	0.00257 3	1	Class 2	Class s 2	Class s 2	Class s 2
37 4	0.88589 4	0.00246 5	1	Class 2	Class s 3	Class s 2	Class s 2
37 5	0.35912 2	0.00379 1	0.93516 3	Class 4	Class s 2	Class s 4	Class s 2
37 6	0.86178 1	0.00201 8	0.99153 4	Class 4	Class s 3	Class s 4	Class s 2
37 7	0.50569 5	0.00469 1	1	Class 2	Class s 2	Class s 2	Class s 2
37 8	0.72549 9	0.00397	0.89707 9	Class 4	Class s 2	Class s 4	Class s 2
37 9	0.79629 1	0.00165 5	0.98108 1	Class 4	Class s 2	Class s 4	Class s 2
38 0	0.10101 6	0.00777 6	0.90770 8	Class 4	Class s 2	Class s 4	Class s 2
38 1	0.52759 2	0.00358 4	0.92166 7	Class 4	Class 2	Class 4	Class 2
38 2	0.25227 4	0.00929 5	0.97310 7	Class 4	Class 2	Class 4	Class 2
38 3	0.91052 5	0.00140 5	1	Class 2	Class 3	Class 2	Class 2
38 4	0.93676 9	0.00130 8	1	Class 2	Class 3	Class 2	Class 2
38 5	0.84227 1	0.00293 5	0.99680 5	Class 4	Class 3	Class 4	Class 2
38 6	0.63111 1	0.00155 1	0.96933 9	Class 4	Class 2	Class 4	Class 2
38 7	0.55615 1	0.00372 1	0.91795 5	Class 4	Class 2	Class 4	Class 2
38 8	0.56441 7	0.00552 7	1	Class 2	Class 2	Class 2	Class 2

38 9	0.15460 1	0.01287	0.87204 7	Class 4	Class 2	Class 4	Class 2
39 0	0.91844 6	0.00211 8	0.96305 3	Class 4	Class s 3	Class s 4	Class s 2
39 1	0.51203 6	0.00509 3	0.96381 6	Class 4	Class s 2	Class s 4	Class s 2
39 2	0.73130 8	0.00343 3	0.92055 6	Class 4	Class s 2	Class s 4	Class s 2
39 3	0.44865 9	0.00662 1	1	Class 2	Class s 2	Class s 2	Class s 2
39 4	0.02838 5	0.07410 9	0.89871 2	Class 4	Class s 2	Class s 4	Class s 2
39 5	0.37927 4	0.00731 5	0.96905 2	Class 4	Class s 2	Class s 4	Class s 2
39 6	0.40642 9	0.00239 5	0.88641 6	Class 4	Class s 2	Class s 4	Class s 2
39 7	0.31467 8	0.00304 5	0.87277 1	Class 4	Class s 2	Class s 4	Class s 2
39 8	0.89128 8	0.00182	0.95233 7	Class 4	Class s 3	Class s 4	Class s 2
39 9	0.78276 5	0.00309 7	0.93680 8	Class 4	Class s 2	Class s 4	Class s 2

40 0	0.27324 8	0.00941 7	0.93641 1	 Class 4	Class 2	Class 4	Class 2
40 1	0.17645 2	0.00565 2	0.90356 2	Class 4	Class 2	Class 4	Class 2
40 2	0.55077 2	0.00828 6	0.84476 3	Class 4	Class 2	Class 4	Class 2
40 3	0.16952 1	0.01649 6	0.96272 5	Class 4	Class 2	Class 4	Class 2
40 4	0.95154 3	0.00307 4	0.99809 2	Class 4	Class 3	Class 4	Class 2
40 5	0.98212 3	0.00142 4	1	Class 2	Class 3	Class 2	Class 2

40 6	0.84089 3	0.00439 4	0.86369 5	Class 4	Clas s 3	Clas s 4	Clas s 2
40 7	0.14819 6	0.01477 6	0.91478 1	Class 4	Clas s 2	Clas s 4	Clas s 2
40 8	0.54276 3	0.02422 6	0.88443 7	Class 4	Clas s 2	Clas s 4	Clas s 2
40 9	0.00857 9	0.36737 4	0.94590 8	Class 4	Clas s 2	Clas s 4	Clas s 2
41 0	0.69216 6	0.00168 1	1	Class 2	Clas s 2	Clas s 2	Clas s 2
41 1	0.22904 7	0.01251 9	0.97737	Class 4	Clas s 2	Clas s 4	Clas s 2
41 2	0.38678 5	0.00943 1	1	Class 2	Clas s 2	Clas s 2	Clas s 2
41 3	0.50543 6	0.00484 7	0.92044 3	Class 4	Clas s 2	Clas s 4	Clas s 2
41 4	0.05518	0.03118 2	0.87297 3	Class 4	Clas s 2	Clas s 4	Clas s 2
41 5	0.7088	0.00408 1	1	Class 2	Clas s 2	Clas s 2	Clas s 2
41 6	0.046	0.01571 9	0.82714 1	Class 4	Clas s 2	Clas s 4	Clas s 2
41 7	0.55148 2	0.00517 4	0.96748 4	Class 4	Clas s 2	Clas s 4	Clas s 2
41 8	0.29659 5	0.00613 9	0.9058	Class 4	Clas s 2	Clas s 4	Clas s 2
41 9	0.63901	0.02077 3	0.82062 6	Class 4	Clas s 2	Clas s 4	Clas s 2
42 0	1	0.00280 2	1	Class 3	Clas s 3	Clas s 2	Clas s 2

421	0.408563	0.034481	0.802272	Class 4	Class 2	Class 4	Class 2
422	0.662579	0.002915	0.907404	Class 4	Class 2	Class 4	Class 2
423	1	0.001024	1	Class 3	Class 3	Class 2	Class 2
424	0.507189	0.003053	0.902483	Class 4	Class 2	Class 4	Class 2
425	0.320207	0.026386	0.826011	Class 4	Class 2	Class 4	Class 2
426	0.470876	0.004456	0.913693	Class 4	Class 2	Class 4	Class 2
427	1	0.001517	0.896051	Class 1	Class 3	Class 4	Class 2
428	0.634814	0.000985	0.983647	Class 4	Class 2	Class 4	Class 2
429	0.175837	0.005581	0.902997	Class 4	Class 2	Class 4	Class 2
430	0.741483	0.00122	0.958075	Class 4	Class 2	Class 4	Class 2
431	0.236912	0.010678	0.852505	Class 4	Class 2	Class 4	Class 2
432	0.808532	0.000775	1	Class 2	Class 3	Class 2	Class 2
433	0.314216	0.006433	0.818634	Class 4	Class 2	Class 4	Class 2
434	0.266599	0.022619	0.730777	Class 4	Class 4	Class 4	Class 4
435	0.869321	0.000918	1	Class 2	Class 3	Class 2	Class 2
436	0.877852	0.000842	1	Class 2	Class 3	Class 2	Class 2
437	0.420615	0.001529	0.796429	Class 4	Class 4	Class 4	Class 4
438	0.934293	0.002542	1	Class 2	Class 3	Class 2	Class 2
439	0.36266	0.003896	0.783238	Class 4	Class 4	Class 4	Class 4
440	0.024276	0.055998	0.752649	Class 4	Class 4	Class 4	Class 4
441	0.614926	0.004233	0.880834	Class 4	Class 2	Class 4	Class 2
442	0.008945	0.279807	0.938082	Class 4	Class 2	Class 4	Class 2
443	0.954979	0.003232	1	Class 2	Class 3	Class 2	Class 2
444	1	0.00148	1	Class 3	Class 3	Class 2	Class 2

KNUST



Appendix C: The Case 2 Results of the DT predictions on the 30% random test dataset

DMU	DEA Class	predictions
1	Class 2	Class 2
2	Class 2	Class 2
3	Class 2	Class 2
4	Class 2	Class 2

5	Class 2	Class 4
6	Class 4	Class 4
7	Class 2	Class 2
8	Class 2	Class 2
9	Class 2	Class 2
10	Class 3	Class 2
11	Class 2	Class 2
12	Class 2	Class 2
13	Class 2	Class 2
14	Class 2	Class 2
15	Class 2	Class 2
16	Class 2	Class 2
17	Class 2	Class 4
18	Class 2	Class 2
19	Class 2	Class 2
20	Class 3	Class 2
21	Class 4	Class 4
22	Class 2	Class 2
23	Class 4	Class 4
24	Class 4	Class 4
25	Class 4	Class 4
26	Class 4	Class 4
27	Class 2	Class 4
28	Class 2	Class 2
29	Class 2	Class 2
30	Class 2	Class 2
31	Class 2	Class 2
32	Class 2	Class 2

33	Class 2	Class 2
----	---------	---------

of the Bank Branches (DMUs)

KNUST



34	Class 2	Class 2
35	Class 2	Class 2
36	Class 2	Class 2
37	Class 2	Class 2
38	Class 2	Class 2
39	Class 2	Class 2
40	Class 2	Class 3

41	Class 2	Class 2
42	Class 2	Class 3
43	Class 2	Class 2
44	Class 3	Class 3
45	Class 2	Class 2
46	Class 2	Class 2
47	Class 2	Class 2
48	Class 2	Class 2
49	Class 2	Class 2
50	Class 2	Class 2
51	Class 2	Class 2
52	Class 2	Class 2
53	Class 2	Class 2
54	Class 2	Class 2
55	Class 4	Class 4
56	Class 2	Class 2
57	Class 4	Class 4
58	Class 4	Class 4
59	Class 4	Class 4
60	Class 4	Class 4
61	Class 4	Class 4
62	Class 4	Class 4
63	Class 4	Class 4
64	Class 4	Class 4
65	Class 4	Class 4

66	Class 2	Class 2
67	Class 2	Class 2
68	Class 2	Class 2

KNUST



69	Class 2	Class 2
70	Class 2	Class 2
71	Class 2	Class 2
72	Class 2	Class 2
73	Class 2	Class 2
74	Class 2	Class 2
75	Class 2	Class 2
76	Class 2	Class 2
77	Class 2	Class 2
78	Class 2	Class 2
79	Class 2	Class 2
80	Class 2	Class 2
81	Class 2	Class 2
82	Class 2	Class 2
83	Class 4	Class 4
84	Class 4	Class 4

85	Class 4	Class 4
86	Class 4	Class 4
87	Class 4	Class 4
88	Class 4	Class 4
89	Class 4	Class 4
90	Class 4	Class 4
91	Class 2	Class 2
92	Class 2	Class 2
93	Class 2	Class 2
94	Class 2	Class 3
95	Class 2	Class 2
96	Class 2	Class 2
97	Class 2	Class 2
98	Class 3	Class 3
99	Class 2	Class 3
100	Class 2	Class 3
101	Class 2	Class 2
102	Class 2	Class 2

KNUST



104	Class 2	Class 2
105	Class 2	Class 2
106	Class 2	Class 2
107	Class 3	Class 3
108	Class 2	Class 2
109	Class 2	Class 2
110	Class 2	Class 2
111	Class 2	Class 2
112	Class 2	Class 2
113	Class 3	Class 3
114	Class 2	Class 2
115	Class 3	Class 3
116	Class 2	Class 2
117	Class 2	Class 2
118	Class 2	Class 2
119	Class 2	Class 2
120	Class 2	Class 2
121	Class 3	Class 3
122	Class 2	Class 3
123	Class 2	Class 2
124	Class 2	Class 2
125	Class 2	Class 2
126	Class 2	Class 2
127	Class 2	Class 2
128	Class 3	Class 3
129	Class 2	Class 2
130	Class 2	Class 2
131	Class 2	Class 3
132	Class 3	Class 3

133	Class 2	Class 2
134	Class 3	Class 3

KNUST



Appendix D: The Case 4 Results of the DT predictions on the 30% dataset of 132 Bank

DMU	DEA Class	predictions
1	Class 2	Class 2

2	Class 2	Class 2
3	Class 2	Class 2
4	Class 2	Class 2
5	Class 2	Class 4
6	Class 4	Class 4
7	Class 2	Class 2
8	Class 2	Class 2
9	Class 2	Class 2
10	Class 2	Class 2
11	Class 2	Class 2
12	Class 2	Class 2
13	Class 2	Class 2
14	Class 2	Class 2
15	Class 2	Class 2
16	Class 2	Class 2
17	Class 2	Class 4
18	Class 2	Class 2
19	Class 2	Class 2
20	Class 2	Class 2
21	Class 4	Class 4
22	Class 2	Class 2
23	Class 4	Class 4
24	Class 4	Class 4
25	Class 4	Class 4
26	Class 4	Class 4
27	Class 2	Class 4
28	Class 2	Class 2
29	Class 2	Class 2

30	Class 2	Class 2
31	Class 2	Class 2
32	Class 2	Class 2
33	Class 2	Class 2

Branches (DMUs)

KNUST



34	Class 2	Class 2
35	Class 2	Class 2
36	Class 2	Class 2
37	Class 2	Class 2
38	Class 2	Class 2
39	Class 2	Class 2
40	Class 2	Class 2

41	Class 2	Class 2
42	Class 2	Class 2
43	Class 2	Class 2
44	Class 2	Class 2
45	Class 2	Class 2
46	Class 2	Class 2
47	Class 2	Class 2
48	Class 2	Class 2
49	Class 2	Class 2
50	Class 2	Class 2
51	Class 2	Class 2
52	Class 2	Class 2
53	Class 2	Class 2
54	Class 2	Class 2
55	Class 4	Class 4
56	Class 2	Class 2
57	Class 4	Class 4
58	Class 4	Class 4
59	Class 4	Class 4
60	Class 4	Class 4
61	Class 4	Class 4
62	Class 4	Class 4
63	Class 4	Class 4
64	Class 4	Class 4
65	Class 4	Class 4

66	Class 2	Class 2
67	Class 2	Class 2
68	Class 2	Class 2

KNUST



69	Class 2	Class 2
70	Class 2	Class 2
71	Class 2	Class 2
72	Class 2	Class 2
73	Class 2	Class 2
74	Class 2	Class 2
75	Class 2	Class 2
76	Class 2	Class 2
77	Class 2	Class 2
78	Class 2	Class 2
79	Class 2	Class 2
80	Class 2	Class 2
81	Class 2	Class 2
82	Class 2	Class 2
83	Class 4	Class 4
84	Class 4	Class 4

85	Class 4	Class 4
86	Class 4	Class 4
87	Class 4	Class 4
88	Class 4	Class 4
89	Class 4	Class 4
90	Class 4	Class 4
91	Class 2	Class 2
92	Class 2	Class 2
93	Class 2	Class 2
94	Class 2	Class 2
95	Class 2	Class 2
96	Class 2	Class 2
97	Class 2	Class 2
98	Class 2	Class 2
99	Class 2	Class 2
100	Class 2	Class 2
101	Class 2	Class 2
102	Class 2	Class 2

KNUST



104	Class 2	Class 2
105	Class 2	Class 2
106	Class 2	Class 2
107	Class 2	Class 2
108	Class 2	Class 2
109	Class 2	Class 2
110	Class 2	Class 2
111	Class 2	Class 2
112	Class 2	Class 2
113	Class 2	Class 2
114	Class 2	Class 2
115	Class 2	Class 2
116	Class 2	Class 2
117	Class 2	Class 2
118	Class 2	Class 2
119	Class 2	Class 2
120	Class 2	Class 2
121	Class 2	Class 2
122	Class 2	Class 2
123	Class 2	Class 2
124	Class 2	Class 2
125	Class 2	Class 2
126	Class 2	Class 2
127	Class 2	Class 2
128	Class 2	Class 2
129	Class 2	Class 2
130	Class 2	Class 2
131	Class 2	Class 2
132	Class 2	Class 2

133	Class 2	Class 2
134	Class 2	Class 2

KNUST



Appendix E: The Case 2 Results of the RF predictions on the 30% random test dataset

DMU	DEA Class	RF Predictions
1	Class 2	Class 2

2	Class 2	Class 2
5	Class 2	Class 2
7	Class 2	Class 2
19	Class 4	Class 4
22	Class 4	Class 4
23	Class 2	Class 2
24	Class 4	Class 4
31	Class 2	Class 2
38	Class 2	Class 2
39	Class 2	Class 3
40	Class 2	Class 2
42	Class 3	Class 3
43	Class 2	Class 2
47	Class 2	Class 2
54	Class 2	Class 2
62	Class 2	Class 2
65	Class 2	Class 4
67	Class 2	Class 4
70	Class 4	Class 4
71	Class 2	Class 2
73	Class 2	Class 2
78	Class 4	Class 4
85	Class 2	Class 2
86	Class 4	Class 4
87	Class 4	Class 4
92	Class 2	Class 2
95	Class 2	Class 2
96	Class 2	Class 2
97	Class 2	Class 2
101	Class 2	Class 2
102	Class 2	Class 2
103	Class 2	Class 2
105	Class 2	Class 2
106	Class 2	Class 2
107	Class 2	Class 2

110	Class 2	Class 2
119	Class 2	Class 2

of the Bank Branches (DMUs)

KNUST



121	Class 2	Class 2
-----	---------	---------

125	Class 2	Class 2
126	Class 2	Class 2
130	Class 2	Class 2
132	Class 3	Class 3
133	Class 2	Class 2
138	Class 2	Class 2
139	Class 2	Class 2
140	Class 2	Class 2
141	Class 2	Class 2
143	Class 2	Class 2
148	Class 3	Class 3
152	Class 2	Class 2
156	Class 2	Class 2
157	Class 2	Class 2
163	Class 2	Class 2
165	Class 2	Class 2
167	Class 2	Class 2
171	Class 2	Class 2
174	Class 4	Class 4
178	Class 2	Class 2
182	Class 4	Class 4
183	Class 4	Class 4
184	Class 4	Class 4
187	Class 4	Class 4
192	Class 4	Class 4
198	Class 4	Class 4
200	Class 4	Class 4
202	Class 4	Class 4
206	Class 2	Class 2
219	Class 2	Class 2
220	Class 2	Class 2
226	Class 2	Class 3
235	Class 2	Class 2
240	Class 2	Class 2

244	Class 2	Class 2
246	Class 2	Class 2
247	Class 2	Class 2
251	Class 2	Class 2
252	Class 2	Class 2

KNUST



254	Class 2	Class 2
260	Class 2	Class 2
268	Class 2	Class 2
273	Class 2	Class 2
275	Class 4	Class 4

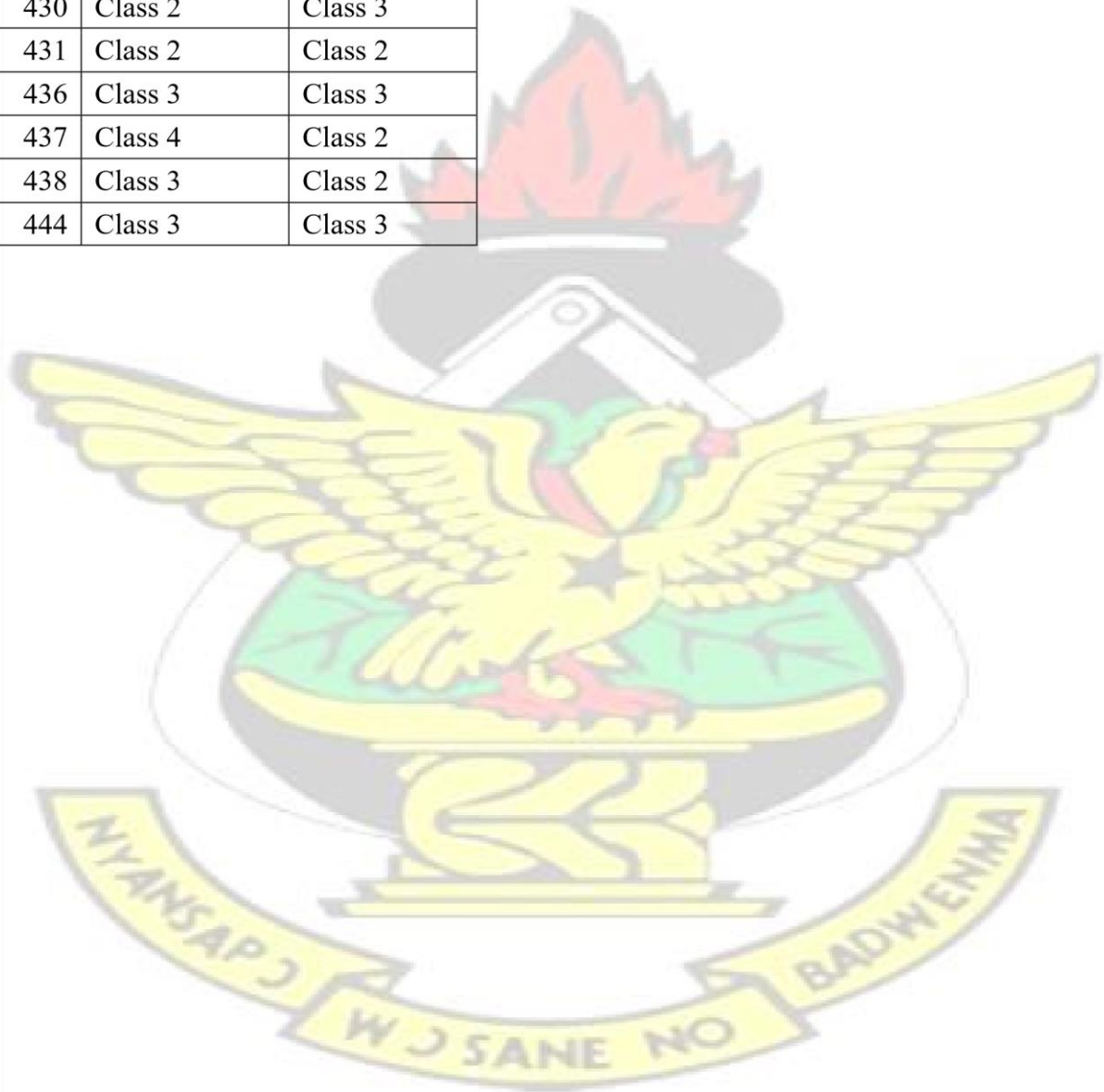
277	Class 4	Class 4
278	Class 4	Class 4
287	Class 4	Class 4
291	Class 4	Class 4
292	Class 4	Class 4
297	Class 2	Class 2
299	Class 2	Class 2
300	Class 2	Class 2
303	Class 2	Class 2
305	Class 2	Class 2
307	Class 2	Class 2
315	Class 2	Class 2
323	Class 2	Class 2
326	Class 2	Class 2
327	Class 2	Class 2
330	Class 3	Class 3
336	Class 2	Class 2
340	Class 3	Class 3
342	Class 2	Class 2
348	Class 3	Class 3
350	Class 2	Class 2
351	Class 2	Class 2
354	Class 2	Class 2
357	Class 3	Class 3
361	Class 2	Class 2
365	Class 2	Class 2
370	Class 2	Class 2
371	Class 2	Class 2
376	Class 3	Class 3
378	Class 2	Class 2

380	Class 2	Class 2
381	Class 2	Class 2
388	Class 2	Class 2
395	Class 2	Class 2
396	Class 2	Class 2

KNUST



405	Class 3	Class 3
406	Class 3	Class 3
408	Class 2	Class 2
410	Class 2	Class 2
413	Class 2	Class 2
414	Class 2	Class 2
415	Class 2	Class 2
422	Class 2	Class 2
426	Class 2	Class 2
429	Class 2	Class 2
430	Class 2	Class 3
431	Class 2	Class 2
436	Class 3	Class 3
437	Class 4	Class 2
438	Class 3	Class 2
444	Class 3	Class 3



Appendix F: The Case 4 Results of the RF predictions on the 30% test dataset of the

DMU	DEA Class	RF Predictions
-----	-----------	----------------

1	Class 2	Class 2
5	Class 2	Class 2
6	Class 2	Class 2
7	Class 2	Class 2
10	Class 2	Class 2
11	Class 2	Class 2
16	Class 4	Class 4
17	Class 2	Class 2
22	Class 4	Class 4
26	Class 4	Class 4
27	Class 2	Class 2
28	Class 2	Class 2
35	Class 2	Class 2
36	Class 2	Class 2
41	Class 2	Class 2
42	Class 2	Class 2
43	Class 2	Class 2
44	Class 2	Class 2
50	Class 2	Class 2
59	Class 2	Class 2
62	Class 2	Class 2
63	Class 2	Class 2
66	Class 2	Class 2
68	Class 4	Class 4
69	Class 2	Class 4
71	Class 2	Class 2
73	Class 2	Class 2
75	Class 2	Class 2
77	Class 4	Class 4
79	Class 2	Class 2
82	Class 4	Class 4
97	Class 2	Class 2
104	Class 2	Class 2
106	Class 2	Class 2
108	Class 2	Class 2

114	Class 2	Class 2
116	Class 2	Class 2
119	Class 2	Class 2

Bank Branches (DMUs)

KNUST



120	Class 2	Class 2
-----	---------	---------

125	Class 2	Class 4
127	Class 2	Class 2
128	Class 2	Class 2
132	Class 2	Class 2
134	Class 2	Class 2
141	Class 2	Class 2
142	Class 2	Class 2
143	Class 2	Class 2
145	Class 2	Class 2
147	Class 2	Class 2
149	Class 2	Class 2
151	Class 2	Class 2
153	Class 2	Class 2
157	Class 2	Class 2
158	Class 2	Class 2
161	Class 2	Class 2
163	Class 2	Class 2
164	Class 2	Class 2
172	Class 2	Class 2
175	Class 4	Class 4
176	Class 4	Class 4
179	Class 4	Class 4
180	Class 4	Class 4
186	Class 4	Class 4
188	Class 4	Class 4
191	Class 4	Class 4
193	Class 4	Class 4
196	Class 4	Class 4
199	Class 4	Class 4
205	Class 2	Class 2
212	Class 2	Class 2
217	Class 2	Class 2
218	Class 2	Class 2
222	Class 2	Class 2

225	Class 2	Class 2
227	Class 2	Class 2
230	Class 2	Class 2
234	Class 2	Class 2
239	Class 2	Class 2

KNUST



242	Class 2	Class 2
246	Class 2	Class 2
263	Class 2	Class 2
264	Class 2	Class 2
265	Class 2	Class 2

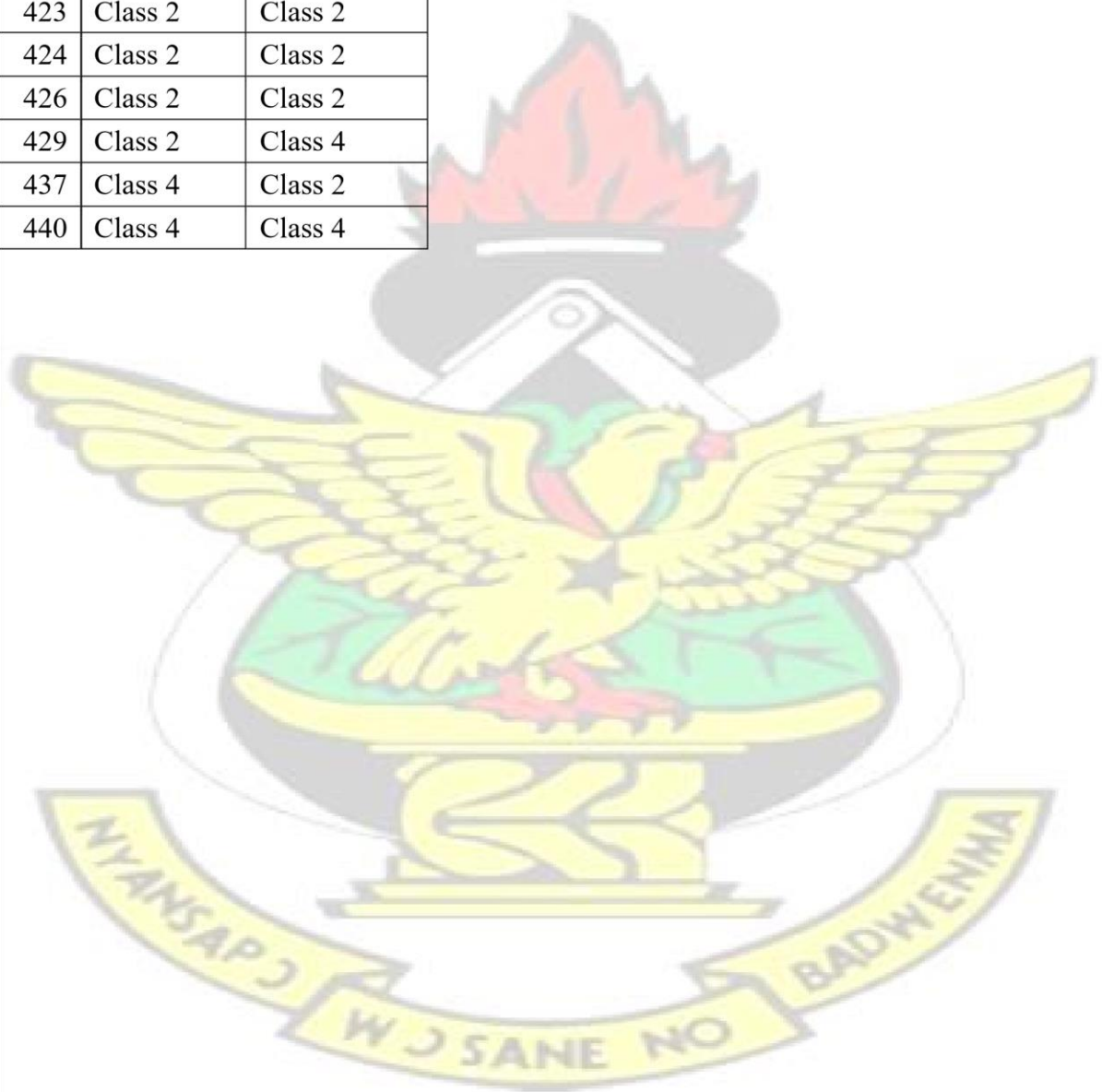
271	Class 2	Class 2
283	Class 4	Class 4
288	Class 4	Class 4
289	Class 4	Class 4
293	Class 2	Class 2
297	Class 2	Class 2
298	Class 2	Class 2
299	Class 2	Class 2
300	Class 2	Class 2
302	Class 2	Class 2
306	Class 2	Class 2
309	Class 2	Class 2
310	Class 2	Class 2
332	Class 2	Class 2
338	Class 2	Class 2
339	Class 2	Class 2
340	Class 2	Class 2
346	Class 2	Class 2
350	Class 2	Class 2
352	Class 2	Class 2
353	Class 2	Class 2
354	Class 2	Class 2
359	Class 2	Class 2
363	Class 2	Class 2
366	Class 2	Class 2
367	Class 2	Class 2
370	Class 2	Class 2
371	Class 2	Class 2
377	Class 2	Class 2
378	Class 2	Class 2

380	Class 2	Class 2
383	Class 2	Class 2
384	Class 2	Class 2
385	Class 2	Class 2
387	Class 2	Class 2

KNUST



390	Class 2	Class 2
391	Class 2	Class 2
397	Class 2	Class 2
400	Class 2	Class 2
401	Class 2	Class 2
404	Class 2	Class 2
407	Class 2	Class 2
412	Class 2	Class 2
414	Class 2	Class 2
415	Class 2	Class 2
423	Class 2	Class 2
424	Class 2	Class 2
426	Class 2	Class 2
429	Class 2	Class 4
437	Class 4	Class 2
440	Class 4	Class 4



DMU	DEA Class	NN Prediction
-----	--------------	------------------

1	2	2.46907306
3	2	2.46907306
9	2	2.46907306
10	2	2.46907306
16	4	2.46907306
17	2	2.46907306
18	3	2.46907306
25	4	2.46907306
26	4	2.987401446
31	2	2.46907306
37	2	2.46907306
40	2	2.46907306
48	2	2.46907306
49	2	2.46907306
52	2	2.46907306
54	2	2.46907306
56	2	2.46907306
59	2	2.46907306
62	2	2.46907306
69	2	2.46907306
77	4	2.46907306
78	4	2.46907306
80	4	2.46907306
85	2	2.46907306
89	4	2.46907306
91	3	2.46907306
92	2	2.298099763
102	2	2.46907306
103	2	2.46907306
104	2	2.46907306
106	2	2.46907306
110	2	2.46907306
116	2	2.46907306
120	2	2.46907306
125	2	2.298099763

128	2	2.46907306	Appendix G: The Case 2 Results of the ANN predictions on the 30% test dataset of the Bank Branches (DMUs)
129	2	2.46907306	
133	2	2.46907306	

KNUST



137	2	2.46907306
-----	---	------------

141	2	2.46907306
145	2	2.46907306
146	3	2.46907306
153	2	2.46907306
156	2	2.46907306
158	2	2.46907306
159	2	2.298099763
160	2	2.46907306
164	2	2.571432751
165	2	2.46907306
172	2	2.298099763
183	4	2.46907306
190	4	2.46907306
191	4	2.46907306
199	4	2.46907306
207	2	2.298099763
210	2	2.46907306
211	2	2.46907306
212	2	2.46907306
215	2	3.0459214
216	2	2.356619717
230	2	2.46907306
231	2	2.46907306
233	2	2.298099763
235	2	2.46907306
236	2	2.46907306
248	2	2.571432751
250	2	2.46907306
251	2	2.46907306
254	2	2.46907306
255	2	2.46907306
256	2	2.46907306
263	2	2.46907306
264	2	2.46907306

270	2	2.46907306
272	2	2.46907306
280	4	2.987401446
284	4	2.46907306
285	4	2.46907306

KNUST



288	4	2.987401446
290	4	2.46907306
303	2	2.46907306
304	2	2.46907306
307	2	2.46907306

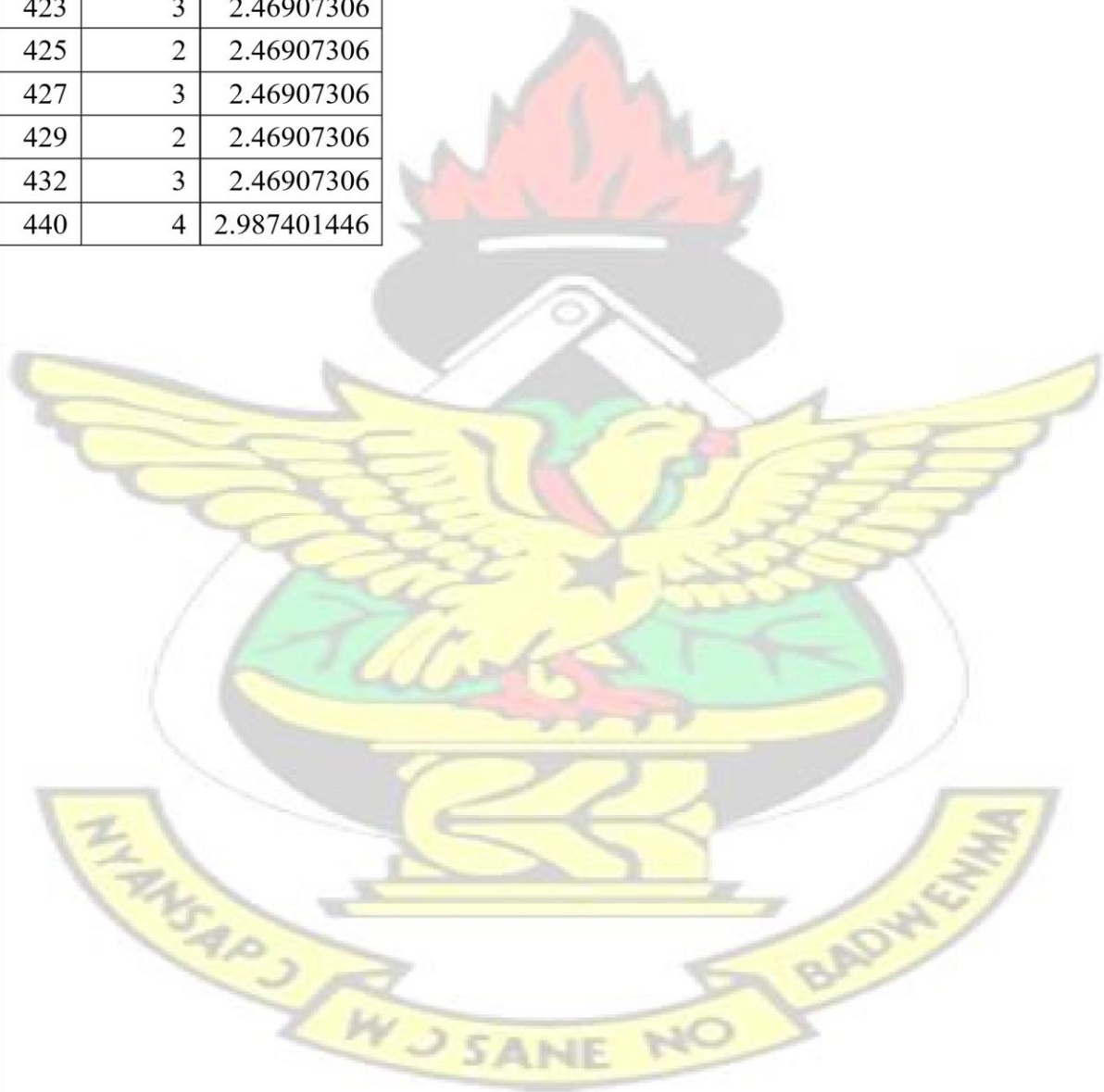
308	2	2.46907306
312	2	2.46907306
313	3	2.46907306
316	2	2.46907306
318	2	2.46907306
320	2	2.46907306
322	2	2.46907306
323	2	2.571432751
326	2	2.46907306
332	2	2.46907306
335	2	2.46907306
338	2	2.46907306
342	2	2.46907306
346	3	2.46907306
347	2	2.46907306
348	3	2.46907306
349	2	2.46907306
350	2	2.46907306
358	2	2.356619717
360	3	2.46907306
362	2	2.46907306
364	2	2.46907306
365	2	2.46907306
366	2	2.46907306
372	2	2.46907306
373	2	2.46907306
378	2	2.46907306
381	2	2.46907306
383	3	2.46907306
384	3	2.46907306

385	3	2.46907306
388	2	2.46907306
394	2	2.987401446
396	2	2.46907306
397	2	2.46907306

KNUST



398	3	2.46907306
400	2	2.46907306
402	2	2.46907306
403	2	2.356619717
408	2	2.46907306
409	2	2.571432751
410	2	2.46907306
414	2	2.987401446
415	2	2.46907306
422	2	2.46907306
423	3	2.46907306
425	2	2.46907306
427	3	2.46907306
429	2	2.46907306
432	3	2.46907306
440	4	2.987401446



Appendix H: The Case 4 Results of the ANN predictions on the 30% test dataset of the

DMU	DEA Class	NN Prediction
-----	-----------	---------------

1	2	2.349055512
3	2	2.349055512
9	2	2.349055512
10	2	2.349055512
16	4	2.349055512
17	2	2.349055512
18	2	2.349055512
25	4	2.349055512
26	4	2.987326403
31	2	2.349055512
37	2	2.349055512
40	2	2.349055512
48	2	2.349055512
49	2	2.349055512
52	2	2.349055512
54	2	2.349055512
56	2	2.349055512
59	2	2.349055512
62	2	2.349055512
69	2	2.349055512
77	4	2.349055512
78	4	2.349055512
80	4	2.349055512
85	2	2.349055512
89	4	2.349055512
91	2	2.349055512
92	2	2.297836691
102	2	2.349055512
103	2	2.349055512
104	2	2.349055512
106	2	2.349055512
110	2	2.349055512
116	2	2.349055512
120	2	2.349055512
125	2	2.297836691

128	2	2.349055512	Bank Branches (DMUs)
129	2	2.349055512	
133	2	2.349055512	

KNUST



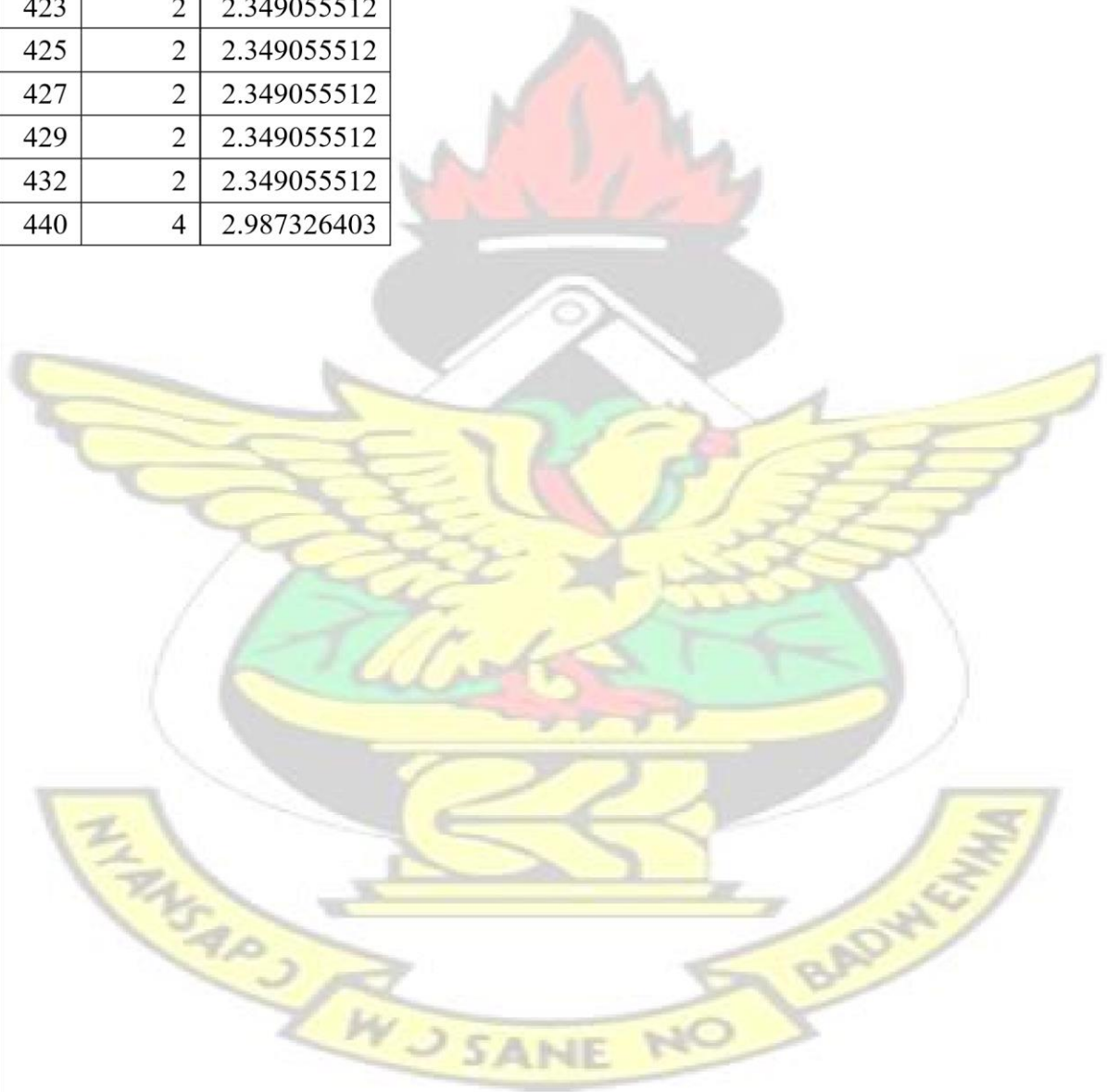
137	2	2.349055512
-----	---	-------------

141	2	2.349055512
145	2	2.349055512
146	2	2.349055512
153	2	2.349055512
156	2	2.349055512
158	2	2.349055512
159	2	2.297836691
160	2	2.349055512
164	2	2.2856747
165	2	2.349055512
172	2	2.297836691
183	4	2.349055512
190	4	2.349055512
191	4	2.349055512
199	4	2.349055512
207	2	2.297836691
210	2	2.349055512
211	2	2.349055512
212	2	2.349055512
215	2	3.046277829
216	2	2.356788118
230	2	2.349055512
231	2	2.349055512
233	2	2.297836691
235	2	2.349055512
236	2	2.349055512
248	2	2.2856747
250	2	2.349055512
251	2	2.349055512
254	2	2.349055512
255	2	2.349055512
256	2	2.349055512
263	2	2.349055512
264	2	2.349055512
270	2	2.349055512
272	2	2.349055512
280	4	2.987326403
284	4	2.349055512
285	4	2.349055512

288	4	2.987326403
290	4	2.349055512
303	2	2.349055512
304	2	2.349055512
307	2	2.349055512

308	2	2.349055512
312	2	2.349055512
313	2	2.349055512
316	2	2.349055512
318	2	2.349055512
320	2	2.349055512
322	2	2.349055512
323	2	2.2856747
326	2	2.349055512
332	2	2.349055512
335	2	2.349055512
338	2	2.349055512
342	2	2.349055512
346	2	2.349055512
347	2	2.349055512
348	2	2.349055512
349	2	2.349055512
350	2	2.349055512
358	2	2.356788118
360	2	2.349055512
362	2	2.349055512
364	2	2.349055512
365	2	2.349055512
366	2	2.349055512
372	2	2.349055512
373	2	2.349055512
378	2	2.349055512
381	2	2.349055512
383	2	2.349055512
384	2	2.349055512
385	2	2.349055512
388	2	2.349055512
394	2	2.987326403
396	2	2.349055512
397	2	2.349055512

398	2	2.349055512
400	2	2.349055512
402	2	2.349055512
403	2	2.356788118
408	2	2.349055512
409	2	2.2856747
410	2	2.349055512
414	2	2.987326403
415	2	2.349055512
422	2	2.349055512
423	2	2.349055512
425	2	2.349055512
427	2	2.349055512
429	2	2.349055512
432	2	2.349055512
440	4	2.987326403



Appendix I: The Results of the DT predictions on the 30% test dataset of the Bank Branches (DMUs)

DMU	ValidSet\$Class	predictions
1	Class 2	Class 2
2	Class 4	Class 2
3	Class 4	Class 2
4	Class 4	Class 2
5	Class 4	Class 2
6	Class 4	Class 2
7	Class 2	Class 2
8	Class 2	Class 2
9	Class 4	Class 2
10	Class 4	Class 2
11	Class 2	Class 2
12	Class 2	Class 2
13	Class 2	Class 2
14	Class 1	Class 1
15	Class 2	Class 4
16	Class 1	Class 2
17	Class 2	Class 2
18	Class 2	Class 2
19	Class 4	Class 2
20	Class 2	Class 4
21	Class 1	Class 4
22	Class 4	Class 4
23	Class 3	Class 2
24	Class 4	Class 2
25	Class 2	Class 2
26	Class 2	Class 2

27	Class 4	Class 2
28	Class 4	Class 1
29	Class 2	Class 4
30	Class 2	Class 1
31	Class 1	Class 2
32	Class 1	Class 4
33	Class 2	Class 2

KNUST



34	Class 3	Class 2	
35	Class 1	Class 4	
36	Class 3	Class 1	
37	Class 2	Class 2	
38	Class 2	Class 2	
39	Class 2	Class 4	
40	Class 3	Class 3	
41	Class 1	Class 3	
42	Class 3	Class 3	
43	Class 3	Class 3	
44	Class 3	Class 3	
45	Class 1	Class 3	
46	Class 1	Class 3	
47	Class 3	Class 4	
48	Class 1	Class 1	
49	Class 3	Class 4	
50	Class 4	Class 4	
51	Class 4	Class 1	
52	Class 2	Class 4	
53	Class 4	Class 4	
54	Class 4	Class 4	
55	Class 4	Class 4	
56	Class 2	Class 4	
57	Class 4	Class 4	
58	Class 4	Class 4	
59	Class 4	Class 4	
60	Class 4	Class 1	
61	Class 4	Class 4	
62	Class 4	Class 1	
63	Class 2	Class 4	
64	Class 4	Class 4	

65	Class 4	Class 4
66	Class 2	Class 4
67	Class 1	Class 4
68	Class 4	Class 4

KNUST



69	Class 4	Class 4			
70	Class 4	Class 4			
71	Class 4	Class 2			
72	Class 4	Class 4			
73	Class 4	Class 4			
74	Class 1	Class 4			
75	Class 4	Class 4			
76	Class 4	Class 4			
77	Class 4	Class 4			
78	Class 4	Class 4			
79	Class 4	Class 4			
80	Class 4	Class 4			
81	Class 1	Class 4			
82	Class 1	Class 4			
83	Class 4	Class 4			
84	Class 4	Class 4			
85	Class 4	Class 4			
86	Class 4	Class 4			
87	Class 4	Class 4			
88	Class 4	Class 4			
89	Class 4	Class 4			
90	Class 4	Class 4			
91	Class 4	Class 4			
92	Class 3	Class 1			
93	Class 3	Class 3			
94	Class 3	Class 2			
95	Class 1	Class 2			
96	Class 1	Class 2			
97	Class 1	Class 2			
98	Class 3	Class 3			
99	Class 3	Class 2			
100	Class 3	Class 3			
101	Class 1	Class 2			
102	Class 1	Class 2			

KNUST



104	Class 3	Class 2
105	Class 1	Class 2
106	Class 3	Class 2
107	Class 3	Class 3
108	Class 3	Class 3
109	Class 1	Class 2
110	Class 3	Class 1
111	Class 3	Class 3
112	Class 2	Class 3
113	Class 3	Class 2
114	Class 2	Class 3
115	Class 2	Class 2
116	Class 2	Class 2
117	Class 2	Class 2
118	Class 4	Class 2
119	Class 4	Class 2
120	Class 4	Class 2
121	Class 2	Class 2
122	Class 2	Class 4
123	Class 4	Class 2
124	Class 2	Class 2
125	Class 4	Class 2
126	Class 2	Class 2
127	Class 2	Class 2
128	Class 2	Class 2
129	Class 2	Class 2
130	Class 4	Class 2
131	Class 2	Class 2
132	Class 2	Class 4

133	Class 2	Class 2
134	Class 2	Class 2

KNUST



Case 2

DMU	ValidSet\$Class	predictions
1	Class 2	Class 3
2	Class 1	Class 4

3	Class 4	Class 4
4	Class 1	Class 4
5	Class 4	Class 4
6	Class 4	Class 2
7	Class 2	Class 2
8	Class 2	Class 2
9	Class 4	Class 4
10	Class 4	Class 1
11	Class 2	Class 3
12	Class 2	Class 1
13	Class 2	Class 3
14	Class 4	Class 4
15	Class 2	Class 4
16	Class 4	Class 1
17	Class 2	Class 2
18	Class 2	Class 1
19	Class 4	Class 2
20	Class 2	Class 4
21	Class 4	Class 4
22	Class 4	Class 4
23	Class 2	Class 2
24	Class 4	Class 2
25	Class 2	Class 2
26	Class 2	Class 2
27	Class 4	Class 2
28	Class 4	Class 4
29	Class 2	Class 4
30	Class 2	Class 4
31	Class 4	Class 3
32	Class 4	Class 4

33	Class 2	Class 3
34	Class 2	Class 2
35	Class 4	Class 4
36	Class 2	Class 4
37	Class 2	Class 3

KNUST



38	Class 2	Class 4
39	Class 2	Class 4
40	Class 2	Class 3
41	Class 1	Class 3
42	Class 3	Class 3
43	Class 3	Class 3

44	Class 2	Class 3
45	Class 1	Class 3
46	Class 1	Class 3
47	Class 2	Class 1
48	Class 1	Class 2
49	Class 2	Class 4
50	Class 4	Class 4
51	Class 4	Class 4
52	Class 2	Class 4
53	Class 1	Class 4
54	Class 1	Class 4
55	Class 4	Class 4
56	Class 2	Class 4
57	Class 4	Class 4
58	Class 1	Class 4
59	Class 1	Class 4
60	Class 1	Class 4
61	Class 1	Class 1
62	Class 4	Class 4
63	Class 2	Class 4
64	Class 4	Class 4
65	Class 4	Class 1
66	Class 3	Class 4
67	Class 1	Class 4
68	Class 4	Class 4
69	Class 1	Class 4
70	Class 1	Class 4
71	Class 4	Class 3

72	Class 1	Class 4
73	Class 1	Class 4

KNUST



74	Class 1	Class 4
75	Class 4	Class 4
76	Class 1	Class 1

77	Class 4	Class 4
78	Class 4	Class 1
79	Class 4	Class 4
80	Class 4	Class 4
81	Class 4	Class 4
82	Class 4	Class 4
83	Class 1	Class 4
84	Class 4	Class 4
85	Class 4	Class 1
86	Class 4	Class 4
87	Class 4	Class 4
88	Class 4	Class 1
89	Class 4	Class 1
90	Class 4	Class 4
91	Class 4	Class 4
92	Class 3	Class 4
93	Class 3	Class 3
94	Class 3	Class 3
95	Class 1	Class 3
96	Class 1	Class 3
97	Class 1	Class 3
98	Class 3	Class 3
99	Class 3	Class 1
100	Class 3	Class 3

101	Class 4	Class 3
102	Class 1	Class 3
103	Class 3	Class 1
104	Class 3	Class 3
105	Class 1	Class 2
106	Class 3	Class 3
107	Class 2	Class 3
108	Class 3	Class 3

KNUST



DMU	ValidSet[, 7]	predValid
1	Class 2	Class 2
3	Class 4	Class 4
6	Class 2	Class 2

8	Class 4	Class 2
109	Class 4	Class 4
110	Class 3	Class 4
111	Class 2	Class 3
112	Class 3	Class 3
113	Class 3	Class 3
114	Class 3	Class 3
115	Class 3	Class 4
116	Class 3	Class 3
117	Class 3	Class 3
118	Class 1	Class 3
119	Class 1	Class 2
120	Class 1	Class 3
121	Class 3	Class 1
122	Class 2	Class 4
123	Class 1	Class 3
124	Class 3	Class 2
125	Class 4	Class 3
126	Class 2	Class 3
127	Class 2	Class 3
128	Class 3	Class 3
129	Class 3	Class 1
130	Class 4	Class 3
131	Class 2	Class 1
132	Class 3	Class 4
133	Class 3	Class 2
134	Class 3	Class 4

**Appendix J: The Results of the RF predictions on the 30% test dataset of the Bank
Branches (DMUs)**

Case 2

9	Class 4	Class 2
12	Class 4	Class 2
15	Class 4	Class 4
17	Class 4	Class 4
18	Class 2	Class 4
19	Class 4	Class 4
28	Class 2	Class 2
37	Class 4	Class 2
38	Class 2	Class 2
44	Class 2	Class 2
46	Class 1	Class 2
47	Class 1	Class 1
49	Class 2	Class 2
50	Class 3	Class 1
56	Class 1	Class 2
58	Class 2	Class 4
59	Class 2	Class 4
60	Class 1	Class 2

62	Class 2	Class 2
65	Class 4	Class 4
68	Class 4	Class 4
70	Class 1	Class 4
71	Class 4	Class 4
75	Class 4	Class 4
79	Class 1	Class 4
88	Class 4	Class 4
95	Class 4	Class 2
97	Class 2	Class 1
99	Class 1	Class 2
100	Class 4	Class 2
103	Class 1	Class 2
108	Class 3	Class 2
114	Class 2	Class 4
119	Class 1	Class 3
120	Class 3	Class 3

123	Class 3	Class 3
-----	---------	---------

KNUST



124	Class 3	Class 3			
126	Class 3	Class 4			
128	Class 3	Class 1			
131	Class 1	Class 3			
132	Class 3	Class 3			
138	Class 1	Class 3			
139	Class 3	Class 3			
142	Class 3	Class 3			
144	Class 3	Class 4			
147	Class 3	Class 4			
149	Class 1	Class 3			
150	Class 3	Class 4			
156	Class 4	Class 4			
157	Class 4	Class 4			
162	Class 4	Class 4			
167	Class 4	Class 4			
169	Class 4	Class 4			
171	Class 2	Class 4			
175	Class 4	Class 4			
180	Class 4	Class 4	208	Class 4	Class 4
183	Class 4	Class 4	216	Class 4	Class 4
187	Class 2	Class 4	219	Class 4	Class 1
190	Class 4	Class 4	220	Class 4	Class 2
193	Class 4	Class 4	221	Class 4	Class 2
204	Class 2	Class 4	228	Class 1	Class 4
207	Class 4	Class 3	230	Class 4	Class 4
			234	Class 4	Class 4
			237	Class 4	Class 4

KNUST



239	Class 4	Class 4
241	Class 4	Class 4
245	Class 4	Class 4
247	Class 4	Class 4
248	Class 4	Class 4
258	Class 4	Class 4

259	Class 4	Class 4
261	Class 4	Class 4
267	Class 4	Class 4
271	Class 4	Class 4
274	Class 4	Class 4
279	Class 4	Class 4
281	Class 4	Class 4
282	Class 4	Class 4
287	Class 4	Class 4
289	Class 4	Class 4
292	Class 4	Class 4
293	Class 4	Class 4
300	Class 4	Class 4
304	Class 4	Class 2
305	Class 4	Class 4
317	Class 3	Class 3
319	Class 3	Class 1
323	Class 1	Class 2
324	Class 3	Class 3
329	Class 3	Class 3
333	Class 1	Class 1
334	Class 3	Class 3
338	Class 3	Class 3
340	Class 3	Class 1
341	Class 1	Class 3
344	Class 1	Class 3
345	Class 3	Class 2

356	Class 3	Class 2
363	Class 1	Class 2
365	Class 3	Class 3

KNUST



367	Class 3	Class 3
370	Class 3	Class 3
372	Class 3	Class 1
377	Class 3	Class 3
379	Class 3	Class 2
381	Class 2	Class 3
385	Class 3	Class 3
388	Class 2	Class 3
389	Class 4	Class 2
390	Class 2	Class 2
392	Class 2	Class 2
396	Class 2	Class 2
403	Class 4	Class 2
409	Class 4	Class 2
410	Class 2	Class 2
411	Class 4	Class 2
413	Class 2	Class 2
417	Class 2	Class 2
418	Class 4	Class 2
419	Class 2	Class 2
428	Class 2	Class 2
435	Class 2	Class 2
437	Class 2	Class 4

Case 4

DMU	ValidSet[, 7]	predValid
1	Class 2	Class 2
3	Class 4	Class 4
6	Class 2	Class 4
8	Class 1	Class 2
9	Class 4	Class 4
12	Class 4	Class 2
15	Class 4	Class 2
17	Class 4	Class 4
18	Class 2	Class 4
19	Class 4	Class 4

28	Class 2	Class 3
37	Class 4	Class 4
38	Class 2	Class 2
44	Class 2	Class 3
46	Class 4	Class 4

47	Class 4	Class 4
49	Class 2	Class 3
50	Class 2	Class 4
56	Class 4	Class 2
58	Class 2	Class 4
59	Class 2	Class 4
60	Class 4	Class 4
62	Class 2	Class 2
65	Class 4	Class 4
68	Class 4	Class 4
70	Class 4	Class 4
71	Class 4	Class 4
75	Class 4	Class 4
79	Class 4	Class 4
88	Class 4	Class 4
95	Class 4	Class 2
97	Class 2	Class 4
99	Class 4	Class 3
100	Class 4	Class 3
103	Class 4	Class 3
108	Class 2	Class 3
114	Class 2	Class 1

119	Class 4	Class 2
120	Class 3	Class 3
123	Class 3	Class 1
124	Class 3	Class 4
126	Class 3	Class 1
128	Class 3	Class 2
131	Class 4	Class 4
132	Class 2	Class 2
138	Class 1	Class 2

KNUST



139	Class 3	Class 2			
142	Class 3	Class 2			
144	Class 2	Class 4			
147	Class 3	Class 4			
149	Class 4	Class 3			
150	Class 3	Class 4			
156	Class 4	Class 4			
157	Class 4	Class 4			
162	Class 4	Class 4			
167	Class 1	Class 4	180	Class 4	Class 4
169	Class 4	Class 4	183	Class 4	Class 4
171	Class 3	Class 4	187	Class 2	Class 1
175	Class 1	Class 4	190	Class 4	Class 1
			193	Class 1	Class 4
			204	Class 3	Class 4
			207	Class 4	Class 4
			208	Class 4	Class 4
			216	Class 1	Class 4
			219	Class 1	Class 4
			220	Class 4	Class 3
			221	Class 4	Class 3
			228	Class 1	Class 4
			230	Class 1	Class 4
			234	Class 1	Class 4
			237	Class 1	Class 4
			239	Class 1	Class 4
			241	Class 4	Class 4
			245	Class 4	Class 4
			247	Class 1	Class 4

248	Class 1	Class 4
258	Class 1	Class 4

KNUST



259	Class 1	Class 4
261	Class 1	Class 4
267	Class 4	Class 4
271	Class 1	Class 4
274	Class 1	Class 4
279	Class 4	Class 1
281	Class 1	Class 4
282	Class 1	Class 4
287	Class 4	Class 4
289	Class 4	Class 4
292	Class 4	Class 4
293	Class 4	Class 4
300	Class 4	Class 4
304	Class 4	Class 4
305	Class 1	Class 4
317	Class 3	Class 3
319	Class 3	Class 3
323	Class 1	Class 3
324	Class 3	Class 3
329	Class 3	Class 3
333	Class 1	Class 4
334	Class 3	Class 3
338	Class 3	Class 3
340	Class 3	Class 4
341	Class 1	Class 2
344	Class 1	Class 3
345	Class 3	Class 3

356	Class 3	Class 3
363	Class 4	Class 2
365	Class 2	Class 3
367	Class 3	Class 2
370	Class 2	Class 3
372	Class 3	Class 4
377	Class 3	Class 3
379	Class 3	Class 3
381	Class 3	Class 2

KNUST



5	4	2.970113
6	2	2.970113
7	2	2.970113
9	4	2.970113
11	4	2.970113
16	2	2.970113

385	Class 3	Class 3
387	Class 3	Class 3
388	Class 3	Class 3
389	Class 1	Class 3
390	Class 3	Class 2
391	Class 3	Class 3
392	Class 3	Class 3
396	Class 3	Class 1
403	Class 1	Class 3
409	Class 1	Class 2
410	Class 2	Class 4
411	Class 1	Class 3
413	Class 3	Class 3
417	Class 3	Class 3
418	Class 1	Class 3
419	Class 2	Class 3
428	Class 2	Class 3
435	Class 3	Class 3
437	Class 2	Class 1

Appendix K: The Results of the ANN predictions on the 30% test dataset of the Bank

Branches (DMUs)

Case 2

DMU	V1	V2
1	2	2.970113
2	4	2.970113
3	4	2.970113
104	1	2.970113
108	3	2.970113
120	3	2.970113

121	3	2.970113
122	1	2.970113
125	1	2.970113
127	3	2.970113
130	1	2.970113
131	1	2.970113
136	3	2.970113
138	1	2.970113
140	3	2.970113

KNUST



141	3	2.970113
146	3	2.970113

150	3	3.220686
157	4	2.970113
161	4	2.970113
162	4	2.970113
163	4	2.970113
164	4	2.970113
165	1	2.970113
168	4	2.970113
169	4	3.220686
171	2	2.970113
172	4	2.970113
174	4	2.970113
179	4	2.970113
180	4	2.970113
184	2	2.970113
186	4	2.970113
190	4	2.970113
193	4	2.970113
196	4	2.970113
197	2	2.970113
201	4	2.970113
202	4	2.970113
210	1	2.970113
212	4	2.970113
216	4	2.970113

218	4	2.970113
226	4	3.220686
228	1	2.970113
229	4	2.970113
235	1	2.970113
237	4	2.970113
238	4	2.970113
239	4	2.970113

KNUST



240	4	2.970113
244	4	2.970113
246	4	2.970113
248	4	2.970113
252	4	2.970113
253	4	2.970113
257	4	2.970113
258	4	2.970113
260	4	2.970113
265	4	2.970113
272	4	2.970113
275	4	2.970113
277	4	2.970113
279	4	2.970113
287	4	2.970113
292	4	2.970113
293	4	2.970113
301	4	2.970113
310	1	2.970113
314	3	2.970113
319	3	2.970113
321	3	2.970113
322	3	2.970113
325	1	2.970113
326	1	2.970113
327	1	2.970113
328	1	2.970113
336	1	2.970113

340	3	2.970113
346	3	2.970113
353	1	2.970113
354	3	2.970113
355	1	2.970113
358	1	2.970113
367	3	2.970113
372	3	2.970113

KNUST

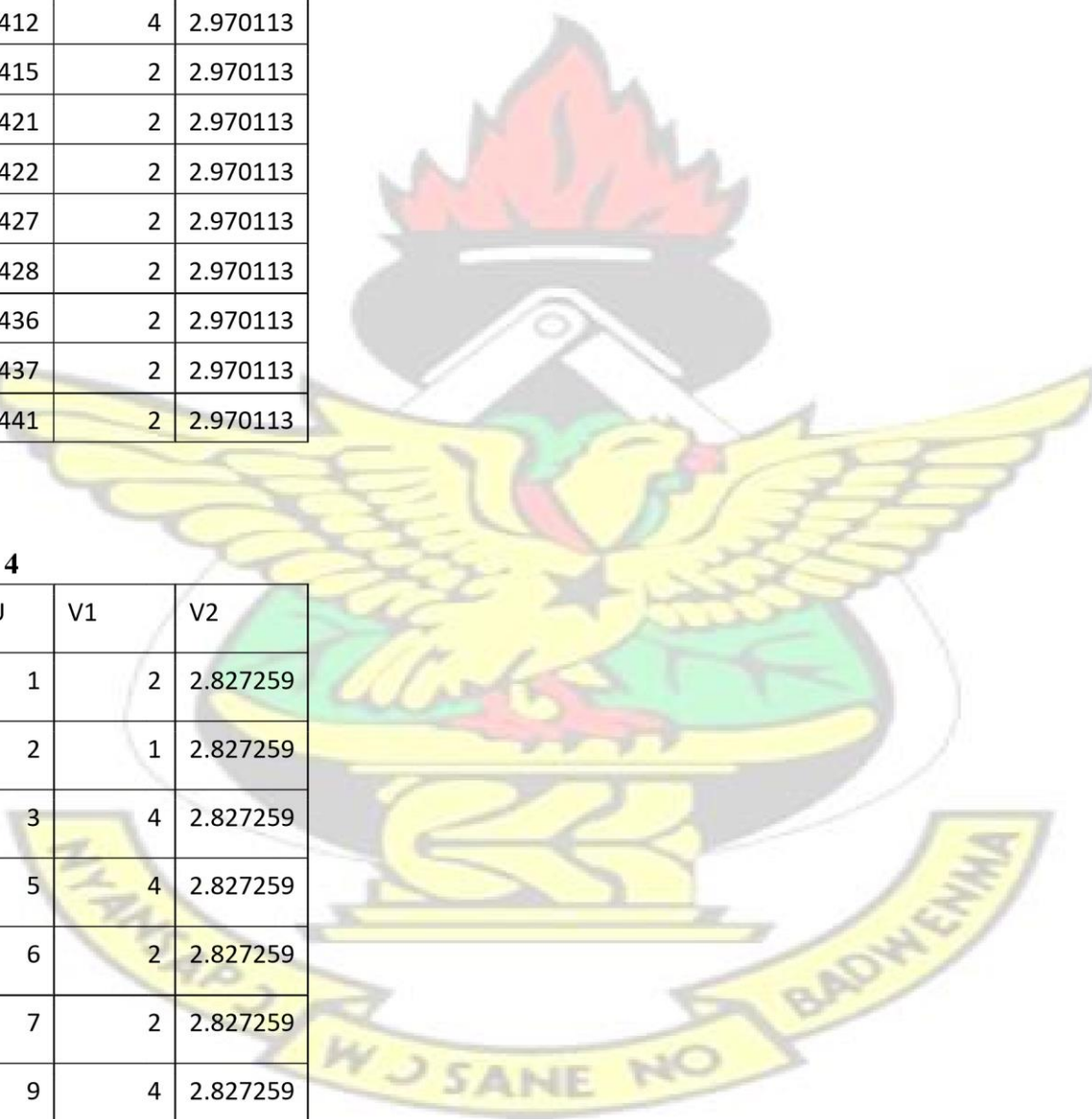


387	2	2.970113
388	2	2.970113
391	2	2.970113
397	4	2.970113
401	4	2.970113
403	4	2.970113
404	2	2.970113
409	4	2.970113
411	4	2.970113
412	4	2.970113
415	2	2.970113
421	2	2.970113
422	2	2.970113
427	2	2.970113
428	2	2.970113
436	2	2.970113
437	2	2.970113
441	2	2.970113

Case 4

DMU	V1	V2
1	2	2.827259
2	1	2.827259
3	4	2.827259
5	4	2.827259
6	2	2.827259
7	2	2.827259
9	4	2.827259
11	4	2.827259
16	2	2.827259
17	4	2.827259

KNUST



19	4	2.827259
21	1	2.827259
23	2	2.827259
29	3	2.827259
34	2	2.827259
37	4	2.827259
41	1	2.827259
45	2	2.827259
54	4	2.827259
56	4	2.827259
57	4	2.827259
66	2	2.827259
75	4	2.827259
76	4	2.827259
78	4	2.827259
81	4	2.827259
82	2	2.827259
84	4	2.827259
87	2	2.827259
91	2	2.827259
97	2	2.827259
99	4	2.827259
103	4	2.827259
104	4	2.827259

108	2	2.827259
120	3	2.827259

KNUST



197	2	2.827259
201	4	2.827259
202	4	2.827259
210	1	2.827259

KNUST



212	4	2.827259			
216	1	2.827259			
218	4	2.827259			
226	4	2.220586			
228	1	2.827259			
229	1	2.827259			
235	4	2.827259			
237	1	2.827259			
238	4	2.827259			
239	1	2.827259			
240	4	2.827259			
244	1	2.827259			
246	4	2.827259			
248	1	2.827259			
			252	4	2.827259
			253	4	2.827259
			257	1	2.827259
			258	1	2.827259
			260	4	2.827259
			265	1	2.827259
			272	4	2.827259
			275	1	2.827259
			277	4	2.827259
			279	4	2.827259
			287	4	2.827259
			292	4	2.827259
			293	4	2.827259
			301	4	2.827259
			310	4	2.827259
			314	3	2.827259
			319	3	2.827259
			321	3	2.827259
			322	2	2.827259
			325	1	2.827259

326	1	2.827259
-----	---	----------

KNUST



327	1	2.827259
328	1	2.827259
336	4	2.827259
340	3	2.827259
346	3	2.827259
353	1	2.827259
354	3	2.827259
355	1	2.827259
358	1	2.827259
367	3	2.827259
372	3	2.827259
387	3	2.827259
388	3	2.827259
391	3	2.827259
397	1	2.827259
401	1	2.827259
403	1	2.827259
404	3	2.827259
409	1	2.827259
411	1	2.827259
412	1	2.827259
415	3	2.827259
421	2	2.827259
422	3	2.827259
427	2	2.827259
428	2	2.827259
436	3	2.827259
437	2	2.827259
441	3	2.827259

KNUST



Appendix L: Prediction by the DEA-LR Case 2

DMU	Predicted	Actual
1	Class 2	Class 2
2	Class 2	Class 2

3	Class 2	Class 2
---	---------	---------

4	Class 2	Class 2
5	Class 2	Class 2
6	Class 2	Class 4
7	Class 2	Class 2
8	Class 2	Class 2
9	Class 2	Class 2
10	Class 2	Class 2
11	Class 2	Class 2
12	Class 2	Class 2
13	Class 2	Class 3
14	Class 2	Class 2
15	Class 2	Class 2
16	Class 2	Class 2
17	Class 2	Class 4
18	Class 2	Class 2
19	Class 2	Class 3
20	Class 4	Class 4
21	Class 4	Class 4
22	Class 4	Class 4
23	Class 4	Class 4
24	Class 4	Class 2
25	Class 2	Class 2
26	Class 2	Class 2
27	Class 2	Class 2
28	Class 2	Class 2
29	Class 2	Class 2
30	Class 2	Class 2
31	Class 2	Class 2

32	Class 2	Class 2
33	Class 2	Class 2
34	Class 2	Class 2
35	Class 2	Class 2
36	Class 2	Class 2
37	Class 2	Class 2
38	Class 2	Class 2

KNUST



39	Class 2	Class 2
40	Class 2	Class 2
41	Class 2	Class 2
42	Class 2	Class 3
43	Class 2	Class 3
44	Class 2	Class 2

45	Class 2	Class 2
46	Class 2	Class 2
47	Class 2	Class 2
48	Class 2	Class 2
49	Class 2	Class 2
50	Class 2	Class 2
51	Class 2	Class 4
52	Class 2	Class 4
53	Class 2	Class 4
54	Class 2	Class 4
55	Class 2	Class 2
56	Class 2	Class 4
57	Class 4	Class 4
58	Class 4	Class 4
59	Class 2	Class 2
60	Class 2	Class 2
61	Class 2	Class 2
62	Class 2	Class 2
63	Class 2	Class 2
64	Class 2	Class 2
65	Class 2	Class 2
66	Class 2	Class 2
67	Class 2	Class 2
68	Class 2	Class 2
69	Class 2	Class 2
70	Class 2	Class 2
71	Class 2	Class 2
72	Class 2	Class 2

73	Class 2	Class 2
74	Class 2	Class 2

KNUST



75	Class 2	Class 2
----	---------	---------

76	Class 2	Class 2
77	Class 2	Class 2
78	Class 2	Class 2
79	Class 2	Class 2
80	Class 2	Class 2
81	Class 4	Class 4
82	Class 4	Class 4
83	Class 4	Class 4
84	Class 4	Class 4
85	Class 4	Class 4
86	Class 4	Class 4
87	Class 4	Class 4
88	Class 4	Class 4
89	Class 4	Class 4
90	Class 2	Class 2
91	Class 2	Class 2
92	Class 2	Class 2
93	Class 2	Class 2
94	Class 2	Class 2
95	Class 2	Class 2
96	Class 2	Class 2
97	Class 2	Class 2
98	Class 2	Class 2
99	Class 2	Class 2
100	Class 2	Class 2
101	Class 2	Class 3

102	Class 2	Class 3
103	Class 2	Class 2
104	Class 2	Class 2
105	Class 2	Class 3
106	Class 2	Class 2
107	Class 2	Class 2
108	Class 2	Class 2
109	Class 2	Class 2

KNUST



110	Class 2	Class 2
111	Class 2	Class 2
112	Class 2	Class 2
113	Class 2	Class 2
114	Class 2	Class 3
115	Class 2	Class 3
116	Class 2	Class 2
117	Class 2	Class 2
118	Class 2	Class 3
119	Class 2	Class 2
120	Class 2	Class 3
121	Class 2	Class 2
122	Class 2	Class 2
123	Class 2	Class 2
124	Class 2	Class 2
125	Class 2	Class 2
126	Class 2	Class 2
127	Class 2	Class 3
128	Class 2	Class 2
129	Class 2	Class 2
130	Class 2	Class 3
131	Class 2	Class 3
132	Class 2	Class 2

Case 4 Prediction by the DEA-LR

DMU	Predicted	Actual
1	Class 2	Class 2
2	Class 2	Class 2
3	Class 2	Class 4
4	Class 2	Class 2
5	Class 2	Class 2
6	Class 2	Class 2
7	Class 2	Class 2
8	Class 2	Class 2
9	Class 2	Class 2
10	Class 2	Class 2

11	Class 2	Class 2
12	Class 2	Class 4
13	Class 2	Class 2
14	Class 4	Class 4
15	Class 2	Class 2
16	Class 4	Class 4
17	Class 4	Class 4
18	Class 4	Class 4
19	Class 4	Class 2
20	Class 2	Class 2
21	Class 2	Class 2
22	Class 2	Class 2
23	Class 2	Class 2
24	Class 2	Class 2
25	Class 2	Class 2
26	Class 2	Class 2

27	Class 2	Class 2
28	Class 2	Class 2
29	Class 2	Class 2
30	Class 2	Class 2
31	Class 2	Class 2
32	Class 2	Class 2
33	Class 2	Class 2
34	Class 2	Class 2
35	Class 2	Class 2
36	Class 2	Class 2
37	Class 2	Class 2
38	Class 2	Class 2
39	Class 2	Class 2
40	Class 2	Class 2
41	Class 2	Class 2
42	Class 2	Class 2
43	Class 2	Class 2
44	Class 2	Class 4

45	Class 2	Class 4
----	---------	---------

KNUST



46	Class 2	Class 4				
47	Class 2	Class 4				
48	Class 2	Class 4				
49	Class 2	Class 4				
50	Class 4	Class 4				
51	Class 2	Class 2				
52	Class 2	Class 2				
53	Class 2	Class 2				
54	Class 2	Class 2				
55	Class 2	Class 2				
56	Class 2	Class 2				
57	Class 2	Class 2				
58	Class 2	Class 2				
59	Class 2	Class 2				
60	Class 2	Class 2				
61	Class 2	Class 2				
62	Class 2	Class 2				
63	Class 2	Class 2				
64	Class 2	Class 2		71	Class 4	Class 4
65	Class 2	Class 2		72	Class 4	Class 4
66	Class 2	Class 2		73	Class 4	Class 4
67	Class 2	Class 2		74	Class 4	Class 4
68	Class 2	Class 2		75	Class 4	Class 4
69	Class 4	Class 4	76	Class 4	Class 4	
70	Class 4	Class 4	77	Class 2	Class 2	
			78	Class 2	Class 2	
			79	Class 2	Class 2	
			80	Class 2	Class 2	

KNUST



81	Class 2	Class 2
82	Class 2	Class 2
83	Class 2	Class 2
84	Class 2	Class 2
85	Class 2	Class 2
86	Class 2	Class 2

87	Class 2	Class 2
88	Class 2	Class 2
89	Class 2	Class 2
90	Class 2	Class 2
91	Class 2	Class 2
92	Class 2	Class 2
93	Class 2	Class 2
94	Class 2	Class 2
95	Class 2	Class 2
96	Class 2	Class 2
97	Class 2	Class 2
98	Class 2	Class 2
99	Class 2	Class 2
100	Class 2	Class 2
101	Class 2	Class 2
102	Class 2	Class 2
103	Class 2	Class 2
104	Class 2	Class 2
105	Class 2	Class 2
106	Class 2	Class 2
107	Class 2	Class 2
108	Class 2	Class 2
109	Class 2	Class 2
110	Class 2	Class 2
111	Class 2	Class 2
112	Class 2	Class 2
113	Class 2	Class 2
114	Class 2	Class 2

115	Class 2	Class 2
116	Class 2	Class 2

KNUST

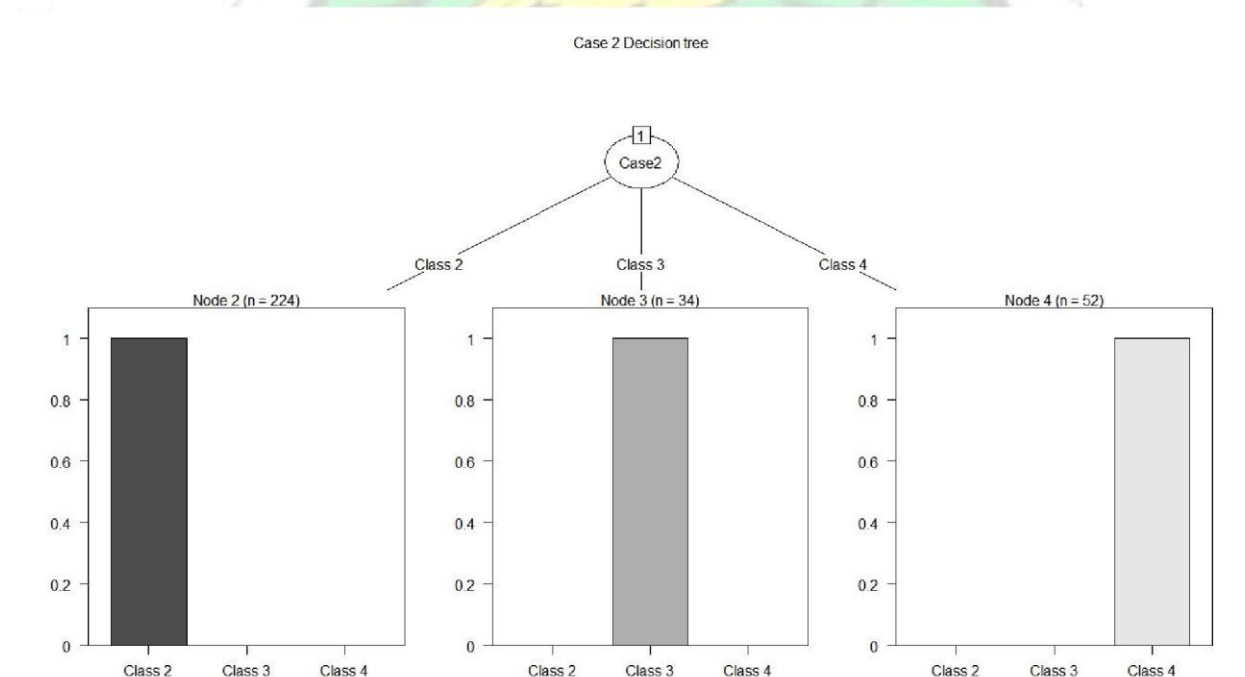


117	Class 2	Class 2
118	Class 2	Class 2
119	Class 2	Class 2
120	Class 2	Class 2
121	Class 2	Class 2
122	Class 2	Class 2
123	Class 2	Class 2
124	Class 2	Class 2
125	Class 2	Class 2
126	Class 2	Class 2
127	Class 2	Class 2
128	Class 2	Class 2
129	Class 2	Class 4
130	Class 2	Class 2
131	Class 2	Class 2
132	Class 2	Class 2

KNUST

Appendix M: The Predictive models Diagrams

Decision Tree Algorithm Predictive model for Case 2



Statistics by Class:

	Class: Class 2	Class: Class 3	Class: Class 4
Sensitivity	1.0000	1.00000	1.0000
Specificity	1.0000	1.00000	1.0000
Pos Pred Value	1.0000	1.00000	1.0000

Neg
Pred

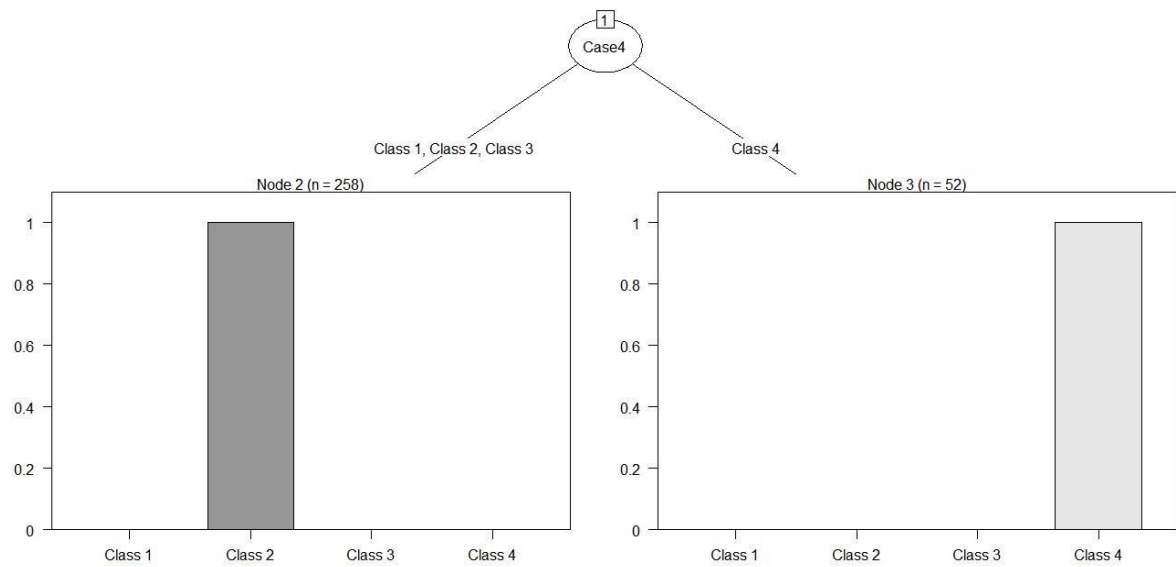
KNUST



Value	1.0000	1.00000	1.0000
Prevalence	0.7313	0.09701	0.1716
Detection Rate	0.7313	0.09701	0.1716
Detection Prevalence	0.7313	0.09701	0.1716
Balanced Accuracy	1.0000	1.00000	1.0000

Decision Tree Algorithm Predictive model for Case 4

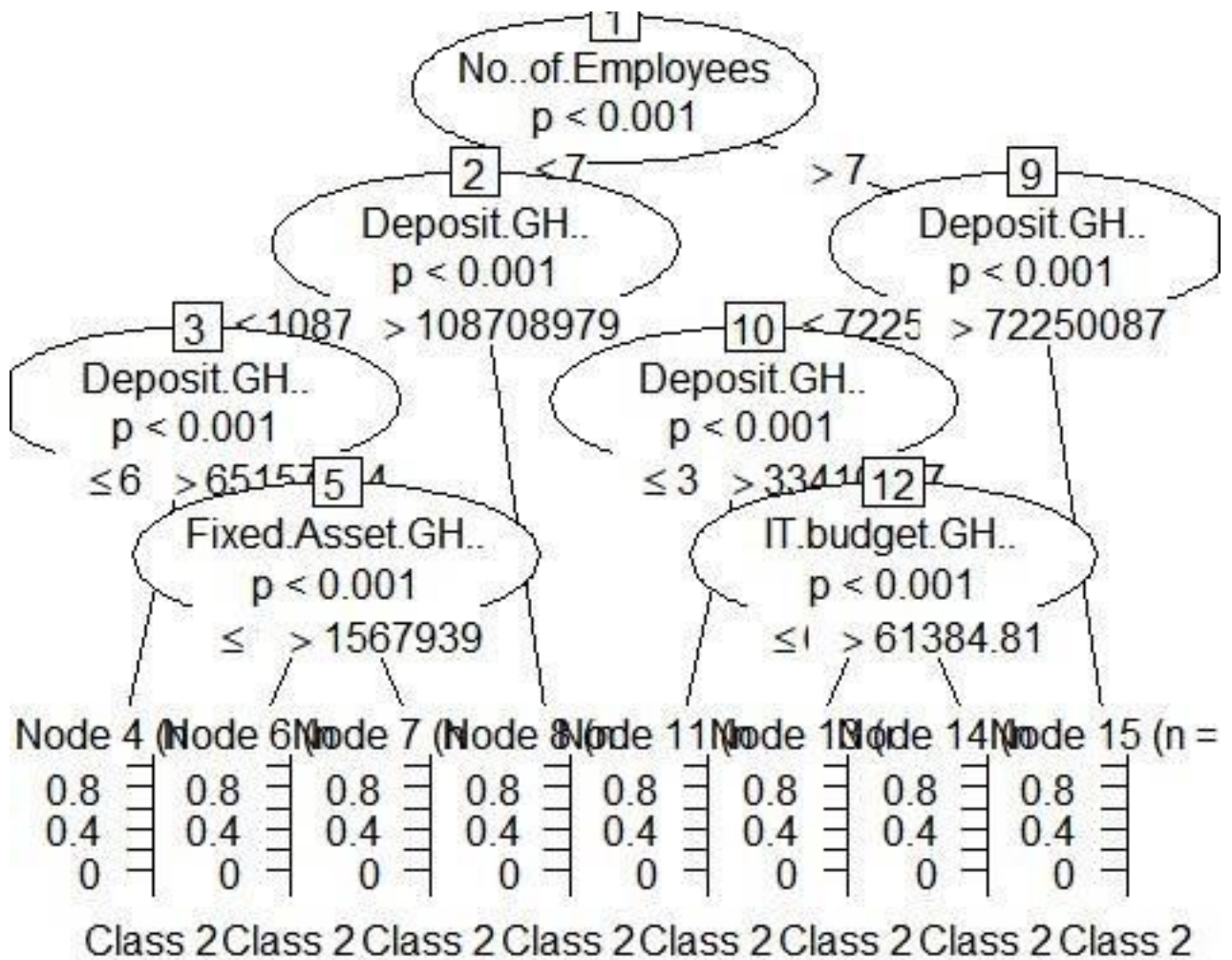
Case 4 Decision tree



Statistics by Class:

	Class: Class 1	Class: Class 2	Class: Class 3	Class: Class 4
Sensitivity	0.000000000	1.00000000	0.000000000	1.00000000
Specificity	1.000000000	0.9166667	1.000000000	1.00000000
Pos Pred Value	NaN	0.9821429	NaN	1.00000000
Neg Pred Value	0.992537313	1.00000000	0.992537313	1.00000000
Prevalence	0.007462687	0.8208955	0.007462687	0.1641791
Detection Rate	0.000000000	0.8208955	0.000000000	0.1641791
Detection Prevalence	0.000000000	0.8358209	0.000000000	0.1641791
Balanced Accuracy	0.500000000	0.9583333	0.500000000	1.00000000

Random Forest Algorithm Predictive model for Case 2

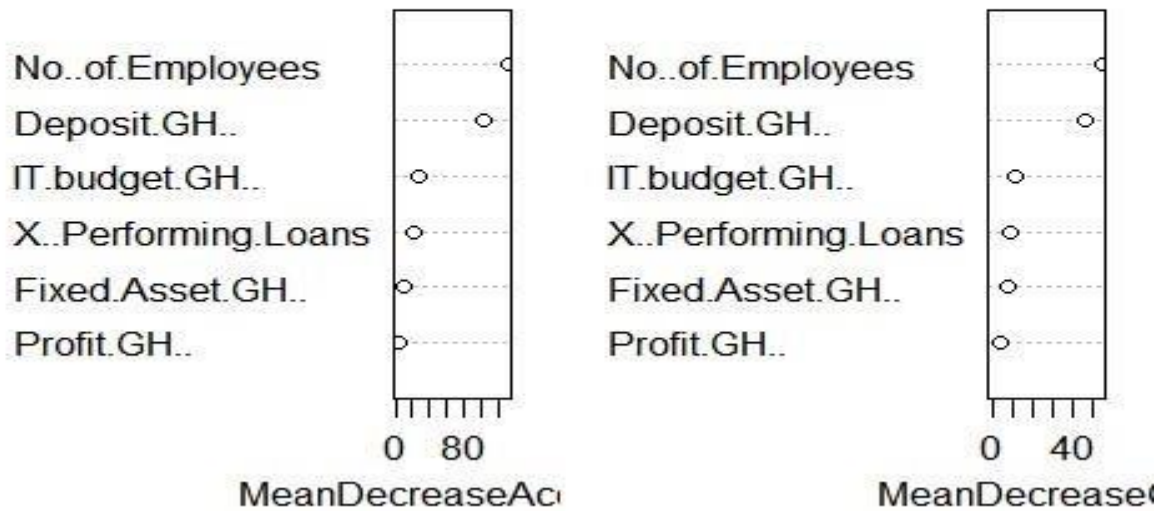


Statistics by Class:

	Class: Class 2	Class: Class 3	Class: Class 4
Sensitivity	0.9490	0.92308	0.9565
Specificity	0.9444	0.97521	0.9820
Pos Pred Value	0.9789	0.80000	0.9167
Neg Pred Value	0.8718	0.99160	0.9909
Prevalence	0.7313	0.09701	0.1716
Detection Rate	0.6940	0.08955	0.1642
Detection Prevalence	0.7090	0.11194	0.1791
Balanced Accuracy	0.9467	0.94914	0.9693

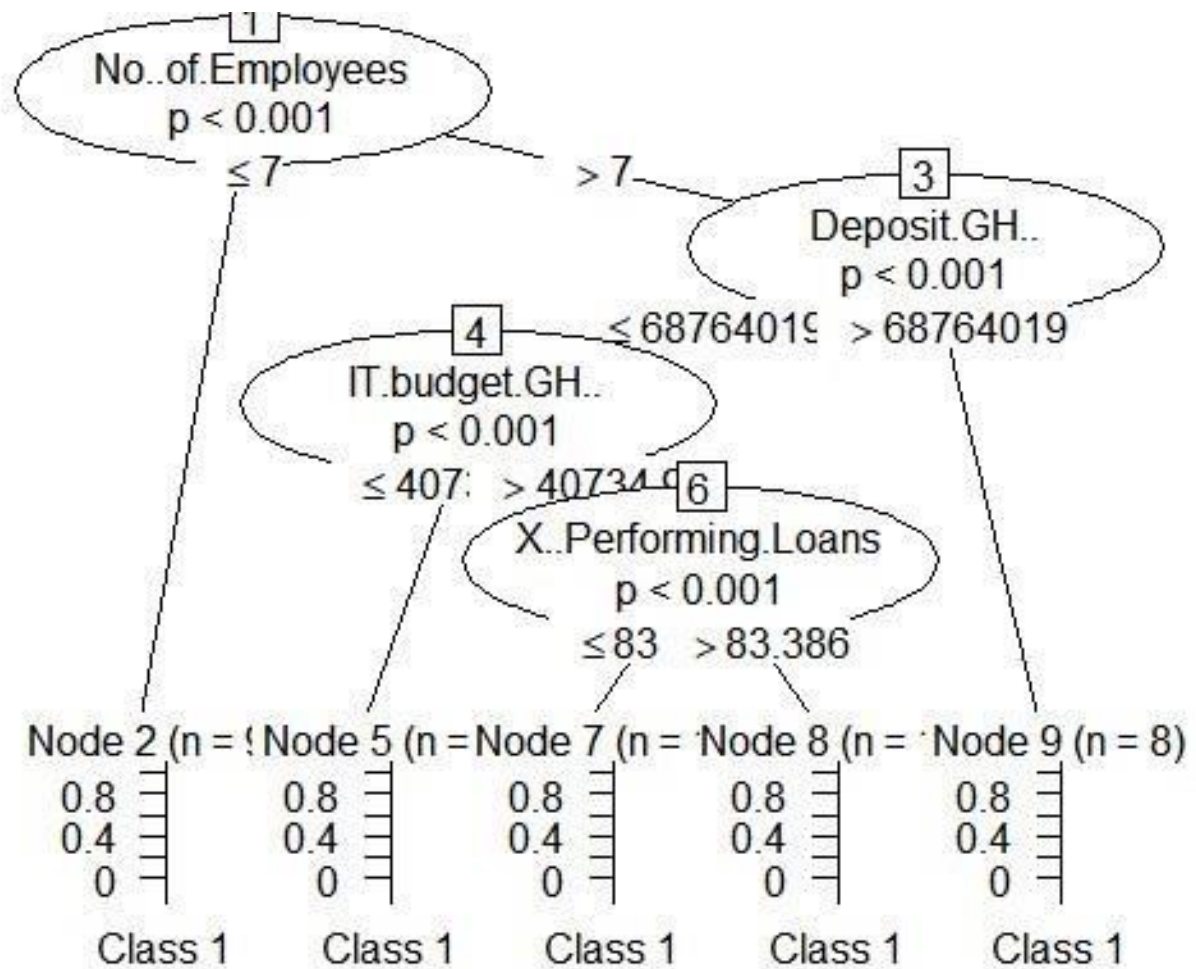
Random Forest Predictive order of significant predictor Variables for Case 2

Case2



Class 2	Class 3	Class 4	MeanDecreaseAccuracy	MeanDecreaseGini
No..of.Employees	63.9262892	8.822786	121.189088	100.808330
X..Performing.Loans	0.3540681	-2.606728	16.168407	13.648501
IT.budget.GH..	8.4985550	9.716305	20.06233	23.214884
Fixed.Asset.GH..	3.5797285	13.145312	3.661803	9.405381
Deposit.GH..	50.1183677	89.151192	27.192690	78.135977
Profit.GH..	1.9453950	-2.994302	12.707079	10.690923

Random Forest Algorithm Predictive model for Case 4

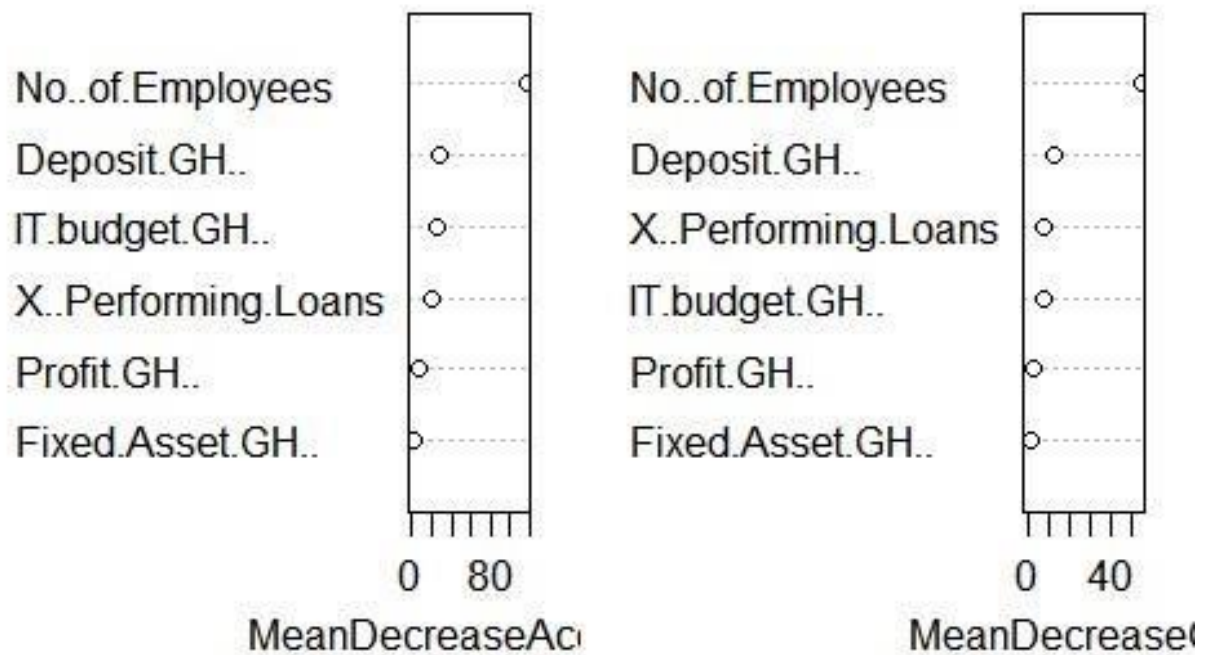


Statistics by Class:

	Class: Class 1	Class: Class 2	Class: Class 3	Class: Class 4
Sensitivity	NA	0.9735	NA	0.9524
Specificity	1	0.9524	1	0.9735
Pos Pred Value	NA	0.9910	NA	0.8696
Neg Pred Value	NA	0.8696	NA	0.9910
Prevalence	0	0.8433	0	0.1567
Detection Rate	0	0.8209	0	0.1493
Detection Prevalence	0	0.8284	0	0.1716
Balanced Accuracy	NA	0.9629	NA	0.9629

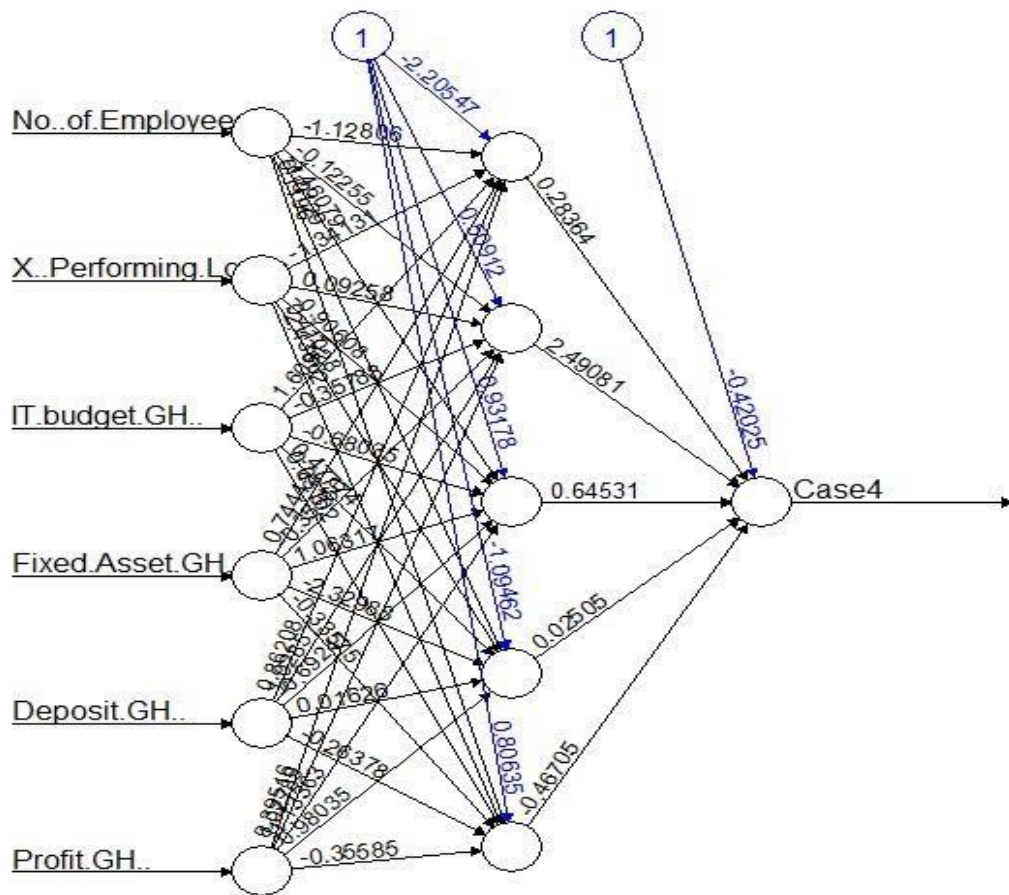
Random Forest Predictive order of significant predictor Variables for Case 4

Case4



Class 1	Class 2	Class 3	Class 4	MeanDecreaseAccuracy
MeanDecreaseGini				
No..of.Employees	0	78.6709688	0	120.4926002
X..Performing.Loans	0	7.9501744	0	15.0642248
IT.budget.GH..	0	19.6747182	0	14.4462853
Fixed.Asset.GH..	0	2.4707794	0	0.6000127
Deposit.GH..	0	19.2456075	0	18.8054602
Profit.GH..	0	0.5260235	0	6.1360222

Artificial Neural Network Predictive model for Case 2



Error: 93.613506 Steps: 35

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.000000000	0.98181818	0.000000000	0.00000000
Specificity	1.000000000	0.08333333	0.969924812	1.00000000
Pos Pred Value	NaN	0.83076923	0.000000000	NaN
Neg Pred Value	0.992537313	0.50000000	0.992307692	0.8358209
Prevalence	0.007462687	0.82089552	0.007462687	0.1641791
Detection Rate	0.000000000	0.80597015	0.000000000	0.00000000
Detection Prevalence	0.000000000	0.97014925	0.029850746	0.00000000
Balanced Accuracy	0.500000000	0.53257576	0.484962406	0.50000000

Appendix N: R codes used for the various analyses

Data Envelopment Analysis Codes

```
install.packages("rDEA") library(rDEA) data1<-  
read.csv(file.choose())  
  
## inputs and outputs for analysis  
  
##Stage 1 Analysis(Input=Employee,IT,Asset:Output=Deposit)
```

```

Y = data1['DEPOSIT']

X = data1[c('Employees', 'IT', 'ASSET')]

## Naive input -oriented DEA score for the first 275 firms under variable returns      -to-scale
firms=1:444

di_stage1 = dea(XREF=X, YREF=Y, X=X[firms,], Y=Y[firms,], model="input",
RTS="constant") di_stage1$thetaOpt write.csv(di_stage1,file = "stage1.csv")

## inputs and outputs for analysis

## Stage 2 Analysis(Input=di_stage1,deposit:Output=Profit,%PL)

Y = data1[c('PROFIT','PL')]

X = data.frame(di_stage1$thetaOpt,data1['DEPOSIT']) firms=1:275

di_stage2 = dea(XREF=X, YREF=Y, X=X[firms,], Y=Y[firms,], model="input",
RTS="variable") di_stage2$thetaOpt write.csv(di_stage2,file = "stage2.csv")

## inputs and outputs for analysis

##Overall Stage Analysis(Input=di_stage2,Employee,IT,Asset::Output=Profit,%PL)

##Y = data1[c('PROFIT','PL')]
##X = data.frame(di_stage2$thetaOpt,data1['Employees','IT','FIXED'])

##firms=1:275

##di_overall = dea(XREF=X, YREF=Y, X=X[firms,], Y=Y[firms,], model="input",
RTS="variable")

##di_overall$thetaOpt

##write.csv(di_overall,file = "overall.csv")

Decision Tree Codes Case 2 library(C50)

library(caret) dataset = read.csv(file.choose(),

header = TRUE)

str(dataset) head(dataset) View(dataset) set.seed(100) train <- sample(nrow(dataset),

0.7*nrow(dataset), replace = FALSE)

```

```
TrainSet <- dataset[train,] ValidSet
```

```
<- dataset[-train,]
```

```
model = C5.0(TrainSet,TrainSet[,7],trials = 10, rules = T, control = C5.0Control(minCases =  
10)) dim(TrainSet)
```

```
dim(ValidSet)
```

```
summary(model)
```

```
test_prediction = predict.C5.0(model, newdata = ValidSet, type ="class") res
```

```
= cbind(ValidSet[,7],as.data.frame(test_prediction))
```

```
confusionMatrix(test_prediction,ValidSet[,7]) write.csv(res,file =  
"predictedCase2.csv")
```

Decision Tree Codes Case 4 library(C50)

```
library(caret) dataset = read.csv(file.choose(),
```

```
header = TRUE)
```

```
str(dataset) head(dataset) View(dataset) set.seed(100) train <-
```

```
sample(nrow(dataset), 0.7*nrow(dataset), replace = FALSE)
```

```
TrainSet <- dataset[train,] ValidSet
```

```
<- dataset[-train,]
```

```
model = C5.0(TrainSet,TrainSet[,7],trials = 10, rules = T, control = C5.0Control(minCases =  
10)) dim(TrainSet)
```

```
dim(ValidSet)
```

```
summary(model)
```

```
test_prediction = predict.C5.0(model, newdata = ValidSet, type ="class") res
```

```
= cbind(ValidSet[,7],as.data.frame(test_prediction))
```



```

confusionMatrix(test_prediction,ValidSet[,7]) write.csv(res,file
= "predictedCase4.csv") Random Forest Codes Case 2
sink("rforest_Case2.doc") library(randomForest)
library(RWeka) library(e1071) library(caret) data1 <-
read.csv(file.choose(), header = TRUE)

head(data1) str(data1)
summary(data1)

# Split into Train and Validation sets
# Training Set : Validation Set = 70 : 30 (random) set.seed(1000)
train <- sample(nrow(data1), 0.7*nrow(data1), replace = FALSE)
TrainSet <- data1[train,] ValidSet
<- data1[-train,]
summary(TrainSet)
summary(ValidSet)

# Create a Random Forest model with default parameters model1 <-
randomForest(Case4 ~ ., data = TrainSet, importance = TRUE) model1
# Fine tuning parameters of Random Forest model

model2 <- randomForest(formula = Case2 ~ ., data = TrainSet, ntree = 500, mtry = 6,
importance = TRUE) model2

# Predicting on train set predTrain <- predict(model1, TrainSet,
type = "class") res =
cbind(data1[train,10],as.data.frame(predTrain))
# Checking classification accuracy table(predTrain, TrainSet$Case2) #

Predicting on Validation set
predValid <-predict(model2, ValidSet, type = "class")
# Checking classification accuracy mean(predValid

```



```
# To check important variables

importance(model2) res =
cbind(ValidSet[,7],as.data.frame(predValid))
write.csv(res,file = "RF_predictedCase4.csv")

varImpPlot(model2) Plotting library(dplyr)
library(gggraph) library(igraph)
install.packages("gggraph") #
install.packages("reptree") tree_func <-
function(final_model,
          tree_num) {
# get tree by index

tree <- randomForest::getTree(final_model,
                              k = tree_num,
                              labelVar = TRUE) %>%
tibble::rownames_to_column() %>%

== ValidSet$Case2)
table(predValid,ValidSet$Case2) confusionMatrix(predValid,
ValidSet$Case2)
```

```

# make leaf split points to NA, so the 0s won't get plotted    mutate(`split
point` = ifelse(is.na(prediction), `split point`, NA))

# prepare data frame for graph    graph_frame <- data.frame(from =
rep(tree$rowname, 2),

                      to = c(tree$`left daughter`, tree$`right daughter`))

# convert to graph and delete the last node that we don't want to plot
graph <- graph_from_data_frame(graph_frame) %>%
delete_vertices("0")

```

```

V(graph)$node_label <- gsub("_", " ", as.character(tree$`split var`))
V(graph)$leaf_label <- as.character(tree$prediction)
V(graph)$split <- as.character(round(tree$`split point`, digits = 2))

# plot

plot <- ggraph(graph, 'dendrogram') + theme_bw() + geom_edge_link() +
geom_node_point() + geom_node_text(aes(label = node_label), na.rm = TRUE, repel =
TRUE) + geom_node_label(aes(label = split), vjust = 2.5, na.rm = TRUE, fill =
"white") + geom_node_label(aes(label = leaf_label, fill = leaf_label), na.rm = TRUE,
repel = TRUE, colour = "white", fontface = "bold", show.legend = FALSE) +
theme(panel.grid.minor = element_blank(), panel.grid.major = element_blank(),

      panel.background = element_blank(),
plot.background = element_rect(fill = "white"),
panel.border = element_blank(), axis.line =
element_blank(), axis.text.x =
element_blank(), axis.text.y =
element_blank(), axis.ticks = element_blank(),
axis.title.x = element_blank(), axis.title.y =
element_blank(), plot.title = element_text(size
= 18)) print(plot)

set node labels
}

# tree_num <- which(model2$finalModel$forest$ndbigtree ==
min(model2$finalModel$forest$ndbigtree)) tree_func(model2,
tree_num = 5)

#plot 2 install.packages("party") library(party)

X = ctree(Case2 ~ ., data=TrainSet, controls=cforest_control(mtry=6, ntree = 500)) plot(X)

```

#

Random Forest Codes Case 4

```
sink("rforest_Case4.doc")
library(randomForest) library(RWeka)
library(e1071) library(caret) data1 <-
read.csv(file.choose(), header = TRUE)
head(data1) str(data1)
summary(data1)
# Split into Train and Validation sets
# Training Set : Validation Set = 70 : 30 (random) set.seed(1000)
train <- sample(nrow(data1), 0.7*nrow(data1), replace = FALSE)
TrainSet <- data1[train,] ValidSet
<- data1[-train,]
summary(TrainSet) summary(ValidSet)
# Create a Random Forest model with default parameters model1 <-
randomForest(Case4 ~ ., data = TrainSet, importance = TRUE) model1
# Fine tuning parameters of Random Forest model
model2 <- randomForest(formula = Case4 ~ ., data = TrainSet, ntree = 500, mtry = 6,
importance = TRUE) model2
# Predicting on train set
predTrain <- predict(model1, TrainSet, type = "class") res
= cbind(data1[train,10],as.data.frame(predTrain))
# Checking classification accuracy table(predTrain,
TrainSet$Case4) # Predicting on Validation set
predValid <-predict(model2, ValidSet, type = "class") #
Checking classification accuracy mean(predValid
```

```
== ValidSet$Case4)
```

```
table(predValid,ValidSet$Case4) confusionMatrix(predValid,  
ValidSet$Case4)
```



```

# To check important variables
importance(model2) res =
cbind(ValidSet[,7],as.data.frame(predValid))
write.csv(res,file = "RF_predictedCase4.csv")
varImpPlot(model2) Plotting library(dplyr)
library(ggraph) library(igraph)
install.packages("ggraph") #
install.packages("reptree") tree_func <-
function(final_model,
         tree_num) { # get tree by
                     tree      <-
randomForest::getTree(final_model,
                     k = tree_num,
                     labelVar = TRUE) %>%
tibble::rownames_to_column() %>%
# make leaf splitpoints to NA, so the 0s won't get plotted mutate(`split
point` = ifelse(is.na(prediction), `split point`, NA))
# prepare data frame for graph graph_frame <-
data.frame(from = rep(tree$rowname, 2),
           to = c(tree$`left daughter`, tree$`right daughter`))
index

```

```

# convert to graph and delete the last node that we don't want to plot
graph <- graph_from_data_frame(graph_frame) %>% delete_vertices("0")

# set node labels
V(graph)$node_label <- gsub("_", " ", as.character(tree$`split var`))

# plot
plot <- ggraph(graph, 'dendrogram') + theme_bw() + geom_edge_link() +
geom_node_point() + geom_node_text(aes(label = node_label), na.rm = TRUE, repel =
TRUE) + geom_node_label(aes(label = split), vjust = 2.5, na.rm = TRUE, fill =
"white") + geom_node_label(aes(label = leaf_label, fill = leaf_label), na.rm = TRUE,
repel = TRUE, colour = "white", fontface = "bold", show.legend = FALSE) +
theme(panel.grid.minor = element_blank(), panel.grid.major = element_blank(),
panel.background = element_blank(),
      plot.background = element_rect(fill = "white"),
panel.border = element_blank(), axis.line =
element_blank(), axis.text.x =
element_blank(), axis.text.y =
element_blank(), axis.ticks = element_blank(),
axis.title.x = element_blank(), axis.title.y =
element_blank(), plot.title = element_text(size
= 18)) print(plot)

# tree_num <- which(model2$finalModel$forest$ndbigtree ==
min(model2$finalModel$forest$ndbigtree)) tree_func(model2,
tree_num = 5)
V(graph)$leaf_label <- as.character(tree$prediction)
V(graph)$split <- as.character(round(tree$`split point`, digits = 2))

```

```
}
```

```
#plot 2 install.packages("party") library(party)
```

```
X = ctree(Case4 ~ ., data=TrainSet, controls=cforest_control(mtry=6, ntree = 500)) plot(X)
```

Artificial Neural Network Codes for Case 2

```

# install.packages('neuralnet')

sink('Case2_nn.doc')

library("neuralnet") library(RWeka)

library(e1071) library(caret)

#Going to create a neural network to perform square rooting #Type
?neuralnet for more information on the neuralnet library data1 <-
read.csv(file.choose(), header = TRUE)

# data1 = data1[,2:8] str(data1)

# this checks for NAs in the data all(is.na.data.frame(data1))

# converting the Case$4 factor type to numeric
unique(data1$Case2) data1$Case2 =
factor(data1$Case2,
        levels = c("Class 2","Class 3",'Class 4'),
labels =c(2,3,4))
data1$Case2 = as.numeric(as.character(data1$Case2))

set.seed(10) train <- sample(nrow(data1), 0.7*nrow(data1),
replace = FALSE)
TrainSet <- data1[train,]
ValidSet <- data1[-train,] # Fitting the neural network nnModel <- neuralnet(Case2 ~
No..of.Employees + X..Performing.Loans + IT.budget.GH.. +
Fixed.Asset.GH.. + Deposit.GH.. + Profit.GH.., data = TrainSet, hidden=5,rep
= 3, threshold = 0.01,err.fct = "sse", act.fct = "logistic",likelihood =
T,linear.output=T)

```

```

print(nnModel) #Plot the
neural network

plot(nnModel)

#testing the neural network on the test dataset

test.results <- compute(nnModel,covariate = ValidSet[,1:6]) #Run them through the neural
network

# calculating the RMSE for the training dataset

predict_testNN = (test.results$net.result * (max(data1$Case2)      - min(data1$Case2))) +
min(data1$Case2)

RMSE.NN = (sum((ValidSet$Case2 - predict_testNN)^2) / nrow(ValidSet)) ^ 0.5

RMSE.NN

# confusioin MAtrix and some self inverted approach

testPred = as.factor(round(test.results$net.result))

testPred =factor(testPred,          levels =
c("2","3","4"),          labels =c(2,3,4))

ValidSetCase2 = as.factor(ValidSet$Case2)

# Confusion Matrix

confusionMatrix(testPred, ValidSetCase2)

# classification Accuracy mean(testPred ==
ValidSet$Case2) #Lets display a better
version of the results test.result =
cbind(ValidSet$Case2, testPred)

# View(test.result) write.csv(test.result, file =
"Case2_nnTestR.csv")

```

Artificial Neural Network Codes for Case 4

```

# install.packages('neuralnet')

sink('Case4_nn.doc')

library("neuralnet") library(RWeka)

library(e1071) library(caret)

```



```

#Going to create a neural network to perform square rooting #Type
?neuralnet for more information on the neuralnet library data1 <-
read.csv(file.choose(), header = TRUE)
# data1 = data1[,2:8] str(data1)
# this checks for NAs in the data all(is.na.data.frame(data1))
# converting the Case$4 factor type to numeric
unique(data1$Case4) data1$Case4 =
factor(data1$Case4,
        levels = c("Class 1", "Class 2", "Class 3", "Class 4"),
labels = c(1,2,3,4))
data1$Case4 = as.numeric(as.character(data1$Case4))
set.seed(10) train <- sample(nrow(data1), 0.7*nrow(data1),
replace = FALSE)
TrainSet <- data1[train,]
ValidSet <- data1[-train,] # Fitting the neural network nnModel <- neuralnet(Case4 ~
No..of.Employees + X..Performing.Loans + IT.budget.GH.. +
Fixed.Asset.GH.. + Deposit.GH.. + Profit.GH..,
data = TrainSet, hidden=5,rep = 3, threshold = 0.01,err.fct = "sse",
act.fct = "logistic",likelihood = T,linear.output=T)
print(nnModel) #Plot
the neural network
plot(nnModel)
#testing the neural network on the test dataset
test.results <- compute(nnModel,covariate = ValidSet[,1:6]) #Run them through the neural
network
# calculating the RMSE for the training dataset
predict_testNN = (test.results$net.result * (max(data1$Case4) - min(data1$Case4))) +
min(data1$Case4)
RMSE.NN = (sum((ValidSet$Case4 - predict_testNN)^2) / nrow(ValidSet)) ^ 0.5
RMSE.NN

```



```
# confusioin MAtrix and some self inverted appproach
testPred = as.factor(round(test.results$net.result)) testPred
=factor(testPred,
```

```
      levels    =    c("2","3"),
labels=c(2,3))
ValidSetCase4 = as.factor(ValidSet$Case4)
```

```
# Confusion Matrix
```

```
confusionMatrix(testPred, ValidSetCase4)
```

```
# classification Accuracy mean(testPred ==
```

```
ValidSet$Case4) #Lets display a better
```

```
version of the results test.result =
```

```
cbind(ValidSet$Case4, testPred)
```

```
# View(test.result) write.csv(test.result, file =
```

```
"Case4_nnTestR.csv") Appendix O: Other
```

Files

```
'data.frame': 444 obs. of  7 variables:
 $ NO..OF.EMPLOYEES: int  7 7 7 7 7 7 7 7 7 7 7 ...
 $ PL              : num  83.8 82.7 83.1 82.5 81.1 ...
 $ IT              : num  60897 13859 34710 23697 47438 ...
 $ ASSET           : num  1454094 1458937 58751 105087 1768646
```

```
 $ DEPOSIT         : num  72757137 17605016 10256530 68830240
30798389 ...
```

```
 $ PROFIT          : num  316872 4372073 1939065 4579781
```

```
...
```

```
778895 ...
```

```
 $ Class           : Factor w/ 4 levels "Class 1","Class 2",...:
2 4 4 2 4 2 2 4 4 2 ...
```

```
=== Summary ===
```

Correctly Classified Instances

217

70

```
%
Incorrectly Classified Instances      93          30
%
Kappa statistic                      0.5523
Mean absolute error                  0.2072
Root mean squared error              0.3218
Relative absolute error              61.1731 %
Root relative squared error          78.2667 %
Total Number of Instances            310
```

=== Confusion Matrix ===

```

  a    b    c    d  <-- classified as   16
4  10  11 |    a = Class 1
    0  45   0  12 |    b = Class 2
    4  13  36  10 |    c = Class 3
    0  29   0 120 |    d = Class 4
```

Confusion Matrix and Statistics

	Reference			
Prediction	Class 1	Class 2	Class 3	Class 4
Class 1	2	1	3	4
Class 2	9	25	8	17
Class 3	3	2	10	0
Class 4	7	10	2	31

Overall Statistics

```

Accuracy : 0.5075
95% CI : (0.4198, 0.5948)
No Information Rate : 0.3881
P-Value [Acc > NIR] : 0.003298
Kappa : 0.2958
McNemar's Test P-Value : 0.023312
```

Statistics by Class:

	Class: Class 1	Class: Class 2	Class: Class 3	Class: Class 4
Sensitivity	0.09524	0.6579	0.43478	0.5962
Specificity	0.92920	0.6458	0.95495	

0.7683			
Pos Pred Value	0.20000	0.4237	0.66667
0.6200			
Neg Pred Value	0.84677	0.8267	0.89076
0.7500			
Prevalence	0.15672	0.2836	0.17164
0.3881			
Detection Rate	0.01493	0.1866	0.07463
0.2313			
Detection Prevalence	0.07463	0.4403	
0.11194	0.3731		
Balanced Accuracy	0.51222	0.6519	
0.69487	0.6822		

'data.frame': 444 obs. of 7 variables:

```

$ NO..OF.EMPLOYEES: int  7 7 7 7 7 7 7 7 7 7 ...
$ PL                : num  83.8 82.7 83.1 82.5 81.1 ...
$ IT                : num  60897 13859 34710 23697 47438 ...
$ ASSET             : num  1454094 1458937 58751 105087 1768646
...
$ DEPOSIT           : num  72757137 17605016 10256530 68830240
30798389 ...
$ PROFIT            : num  316872 4372073 1939065 4579781 778895
...
$ Class             : Factor w/ 4 levels "Class 1","Class 2",...:
2 1 4 3 4 2 2 1 4 2 ...

```

=== Summary ===

Correctly Classified Instances	201	64.8387
%		
Incorrectly Classified Instances	109	35.1613
%		
Kappa statistic	0.4894	
Mean absolute error	0.2426	
Root mean squared error	0.3483	Relative
absolute error	68.0854 %	
Root relative squared error	82.5415 %	
Total Number of Instances	310	

=== Confusion Matrix ===

```

a    b    c    d    <-- classified as    22
1      13   26 |    a = Class 1
2      24   14    9 |    b = Class 2
3      9   45   14 |    c = Class 3    1    7    10 110 |    d =
Class 4

```

Confusion Matrix and Statistics

Reference				
Prediction	Class 1	Class 2	Class 3	Class 4
Class 1	2	4	4	7
Class 2	3	7	2	4
Class 3	10	11	15	5
Class 4	14	12	6	28

Overall Statistics

```

Accuracy : 0.3881
95% CI : (0.3052, 0.476)
No Information Rate : 0.3284
P-Value [Acc > NIR] : 0.08513
Kappa : 0.1658
Mcnemar's Test P-Value : 0.01757

```

Statistics by Class:

	Class: Class 1	Class: Class 2	Class: Class 3
3 Class: Class 4			
Sensitivity	0.06897	0.20588	0.5556
0.6364			
Specificity	0.85714	0.91000	0.7570
0.6444			
Pos Pred Value	0.11765	0.43750	0.3659
0.4667			
Neg Pred Value	0.76923	0.77119	
0.8710			
Prevalence	0.21642	0.25373	0.2015
0.3284			
Detection Rate	0.01493	0.05224	0.1119
0.2090			
Detection Prevalence	0.12687	0.11940	0.3060
0.4478			
Balanced Accuracy	0.46305	0.55794	

0.6563

0.6404

```
'data.frame': 444 obs. of 7 variables:
 $ NO..OF.EMPLOYEES: int  7 7 7 7 7 7 7 7 7 7 7 ...
 $ PL               : num  83.8 82.7 83.1 82.5 81.1 ...
 $ IT               : num  60897 13859 34710 23697 47438 ...
 $ ASSET            : num  1454094 1458937 58751 105087 1768646
 ...
 $ DEPOSIT          : num  72757137 17605016 10256530 68830240
30798389 ...
 $ PROFIT           : num  316872 4372073 1939065 4579781 778895
 ...
 $ Class            : Factor w/ 4 levels "Class 1","Class 2",...:
2 1 4 3 4 2 2 1 4 2 ...
[1] Class 2 Class 1 Class 4 Class 3
Levels: Class 1 Class 2 Class 3 Class 4
Confusion Matrix and Statistics
```

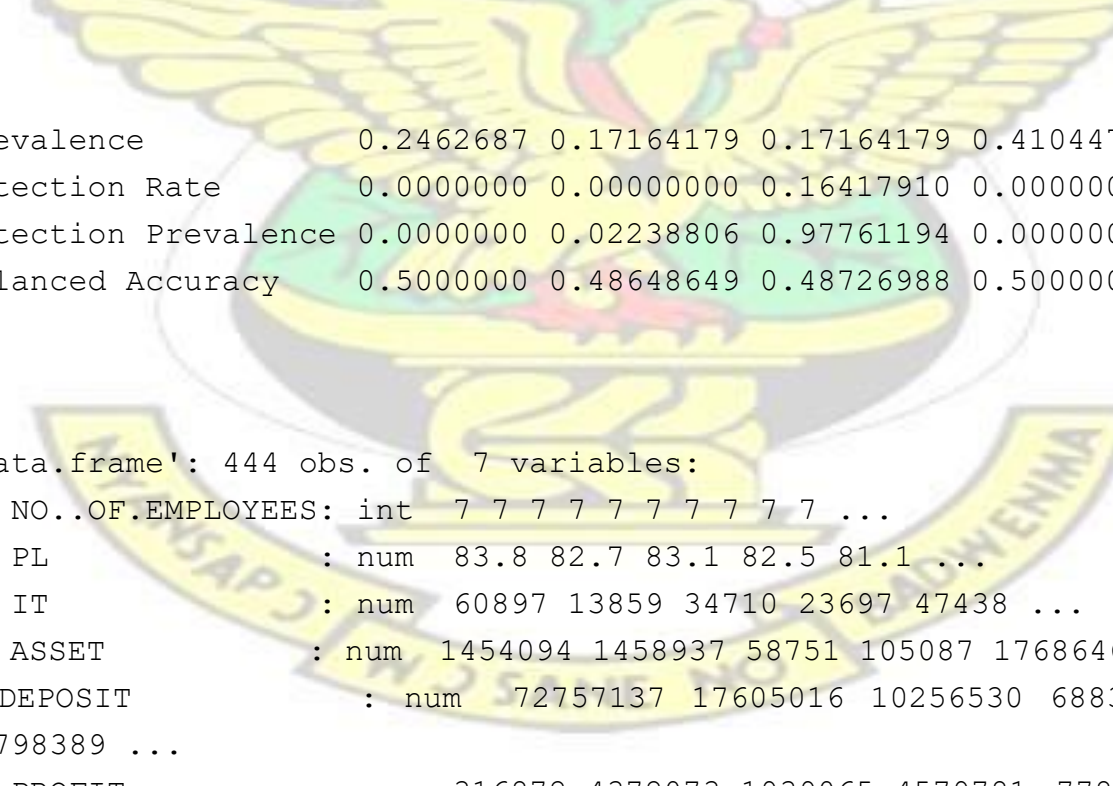
	Reference			
Prediction	1	2	3	4
1	0	0	0	0
2	0	0	1	2
3	33	23	22	53
4	0	0	0	0

Overall Statistics

```
Accuracy : 0.1641791
95% CI : (0.1058435, 0.2379513)
No Information Rate : 0.4104478
P-Value [Acc > NIR] : 1
Kappa : -0.009009
Mcnemar's Test P-Value : NA
```

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.0000000	0.0000000	0.95652174	0.0000000
Specificity	1.0000000	0.97297297	0.01801802	1.0000000
Pos Pred Value	NaN	0.00000000	0.16793893	NaN
Neg Pred Value	0.7537313	0.82442748	0.66666667	0.5895522



Prevalence 0.2462687 0.17164179 0.17164179 0.4104478
Detection Rate 0.0000000 0.00000000 0.16417910 0.00000000
Detection Prevalence 0.0000000 0.02238806 0.97761194 0.00000000
Balanced Accuracy 0.5000000 0.48648649 0.48726988 0.50000000

'data.frame': 444 obs. of 7 variables:
 \$ NO..OF.EMPLOYEES: int 7 7 7 7 7 7 7 7 7 7 ...
 \$ PL : num 83.8 82.7 83.1 82.5 81.1 ...
 \$ IT : num 60897 13859 34710 23697 47438 ...
 \$ ASSET : num 1454094 1458937 58751 105087 1768646 ...
 \$ DEPOSIT : num 72757137 17605016 10256530 68830240
 30798389 ...
 \$ PROFIT : num 316872 4372073 1939065 4579781 778895
 ...
 \$ Class : Factor w/ 4 levels "Class 1","Class 2",...:
 2 4 4 2 4 2 2 4 4 2 ...
[1] FALSE
[1] Class 2 Class 4 Class 3 Class 1
Levels: Class 1 Class 2 Class 3 Class 4
Confusion Matrix and Statistics

	Reference			
Prediction	1	2	3	4
1	0	0	0	0
2	0	0	0	0
3	24	27	19	64
4	0	0	0	0

Overall Statistics

Accuracy : 0.141791
95% CI : (0.0875778, 0.2125375)
No Information Rate : 0.4776119
P-Value [Acc > NIR] : 1
Kappa : 0
Mcnemar's Test P-Value : NA

Statistics by Class:



	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	0.0000000	0.0000000	1.0000000	0.0000000
Specificity	1.0000000	1.0000000	0.0000000	1.0000000
Pos Pred Value	NaN	NaN	0.141791	NaN
Neg Pred Value	0.8208955	0.7985075	NaN	0.5223881
Prevalence	0.1791045	0.2014925	0.141791	0.4776119
Detection Rate	0.0000000	0.0000000	0.141791	0.0000000
Detection Prevalence	0.0000000	0.0000000	1.0000000	0.0000000
Balanced Accuracy	0.5000000	0.5000000	0.5000000	0.5000000

NO..OF.EMPLOYEES	PL	IT	ASSET	DEPOSIT
PROFIT	Class			
1	7	83.75370	60897.36	1454093.72
316871.7	Class 2			
2	7	82.69211	13859.08	1458937.49
				17605016
				4372073.0
	Class 4			
3	7	83.11167	34710.15	58751.02
				10256530
				1939064.7
	Class 4			
4	7	82.54663	23696.85	105086.50
				68830240
				4579780.7
	Class 2			
5	7	81.09985	47438.36	1768645.92
778894.6	Class 4			
6	7	83.27801	27897.11	231642.89
1098903.5	Class 2			

Confusion Matrix and Statistics

Reference				
Prediction	Class 1	Class 2	Class 3	Class 4
Class 1	2	1	5	1
Class 2	7	16	4	14
Class 3	6	2	16	1
Class 4	3	8	4	44

Overall Statistics

Accuracy : 0.5821
 95% CI : (0.4938, 0.6667)
 No Information Rate : 0.4478

P-Value [Acc > NIR] : 0.001219

Kappa : 0.3959

Mcnemar's Test P-Value : 0.138147

Statistics by Class:

	Class: Class 1	Class: Class 2	Class: Class
3 Class: Class 4			
Sensitivity	0.11111	0.5926	0.5517
0.7333			
Specificity	0.93966	0.7664	0.9143
0.7973			
Pos Pred Value	0.22222	0.3902	0.6400
0.7458			
Neg Pred Value	0.87200	0.8817	
0.8807	0.7867		
Prevalence	0.13433	0.2015	0.2164
0.4478			
Detection Rate	0.01493	0.1194	0.1194
0.3284			
Detection Prevalence	0.06716	0.3060	0.1866
0.4403			
Balanced Accuracy	0.52538	0.6795	
0.7330	0.7653		
MeanDecreaseGini			
NO..OF.EMPLOYEES	14.25971		
IT	57.58200		
ASSET	52.62596		

NO..OF.EMPLOYEES	PL	IT	ASSET	DEPOSIT	PROFIT
Class					
1	7 83.75370	60897.36	1454093.72	72757137	
316871.7 Class 2					
2	7 82.69211	13859.08	1458937.49	17605016	
4372073.0					
Class 1					
3	7 83.11167	34710.15	58751.02	10256530	
1939064.7					
Class 4					
4	7 82.54663	23696.85	105086.50	68830240	
4579780.7 Class 3					

5 7 81.09985 47438.36 1768645.92 30798389
 778894.6 Class 4
 6 7 83.27801 27897.11 231642.89 52158533
 1098903.5 Class 2

Confusion Matrix and Statistics

	Reference			
Prediction	Class 1	Class 2	Class 3	Class 4
Class 1	0	3	3	2
Class 2	4	4	6	6
Class 3	6	8	17	6
Class 4	21	8	7	33

Overall Statistics

Accuracy : 0.403
 95% CI : (0.3192, 0.4911)
 No Information Rate : 0.3507
 P-Value [Acc > NIR] : 0.120256
 Kappa : 0.1615
 Mcnemar's Test P-Value : 0.007651

Statistics by Class:

	Class: Class 1	Class: Class 2	Class: Class 3	Class: Class 4
Sensitivity	0.0000	0.17391		
0.5152	0.7021			
Specificity	0.9223	0.85586		0.8020
0.5862				
Pos Pred Value	0.0000	0.20000		0.4595
0.4783				
Neg Pred Value	0.7540	0.83333		0.8351
0.7846				
Prevalence	0.2313	0.17164		
0.2463	0.3507			
Detection Rate	0.0000	0.02985		0.1269
0.2463				
Detection Prevalence	0.0597	0.14925		0.2761
0.5149				
Balanced Accuracy	0.4612	0.51488		
0.6586	0.6442			
MeanDecreaseGini				

```
NO..OF.EMPLOYEES      17.13943
IT                     59.38962
ASSET                  56.06401
```

'data.frame': 444 obs. of 7 variables:

```
$ NO..OF.EMPLOYEES: int  7 7 7 7 7 7 7 7 7 7 ...
$ PL               : num  83.8 82.7 83.1 82.5 81.1 ...
$ IT               : num  60897 13859 34710 23697 47438 ...
$ ASSET            : num  1454094 1458937 58751 105087 1768646
...
$ DEPOSIT          : num  72757137 17605016 10256530 68830240
30798389 ...
$ PROFIT           : num  316872 4372073 1939065 4579781 778895
...
$ Case.2           : Factor w/ 3 levels "Class 2","Class 3",...:
1 1 1 2 1 1 1 1 1 1 ...
```

=== Summary ===

```
Correctly Classified Instances      299          96.4516
%
Incorrectly Classified Instances     11          3.5484
%
Kappa statistic                     0.9204
Mean absolute error                  0.0428
Root mean squared error              0.1463
Relative absolute error              14.4526 %
Root relative squared error          38.1082 %
Total Number of Instances           310
```

=== Confusion Matrix ===

```

a   b   c   <-- classified as 217
4   2 |   a = Class 2
    2 33   1 |   b = Class 3
    2   0 49 |   c = Class 4
```

Confusion Matrix and Statistics

```

              Reference
Prediction Class 2 Class 3 Class 4
Class 2      89      2      0
Class 3       7      9      0
```

Class 4 3 0 24

Overall Statistics

Accuracy : 0.9104
95% CI : (0.8488, 0.9529)
No Information Rate : 0.7388
P-Value [Acc > NIR] : 5.296e-07
Kappa : 0.802
Mcnemar's Test P-Value : NA

Statistics by Class:

	Class: Class 2	Class: Class 3	Class: Class 4
Sensitivity	0.8990	0.81818	1.0000
Specificity	0.9429	0.94309	0.9727
Pos Pred Value	0.9780	0.56250	0.8889
Neg Pred Value	0.7674	0.98305	1.0000
Prevalence	0.7388	0.08209	0.1791
Detection Rate	0.6642	0.06716	0.1791
Detection Prevalence	0.6791	0.11940	0.2015
Balanced Accuracy	0.9209	0.88064	0.9864

'data.frame': 444 obs. of 7 variables:
\$ NO..OF.EMPLOYEES: int 7 7 7 7 7 7 7 7 7 7 7 ...
\$ PL : num 83.8 82.7 83.1 82.5 81.1 ...
\$ IT : num 60897 13859 34710 23697 47438 ...
\$ ASSET : num 1454094 1458937 58751 105087 1768646 ...
\$ DEPOSIT : num 72757137 17605016 10256530 68830240 30798389 ...
\$ PROFIT : num 316872 4372073 1939065 4579781 778895 ...
\$ Class : Factor w/ 4 levels "Class 1","Class 2",...:
2 2 2 2 2 2 2 2 2 2 ...

=== Summary ===

Correctly Classified Instances	303	97.7419
%		

Incorrectly Classified Instances	7	2.2581
%		
Kappa statistic	0.9194	
Mean absolute error	0.0209	
Root mean squared error	0.1023	
Relative absolute error	14.5603 %	
Root relative squared error	38.5556 %	
Total Number of Instances	310	

=== Confusion Matrix ===

	a	b	c	d	<-- classified as	0
0	0	1			a = Class 1	
	0	255	0	3	b = Class 2	
	0	1	0	0	c = Class 3	
	0	2	0	48	d = Class 4	

Confusion Matrix and Statistics

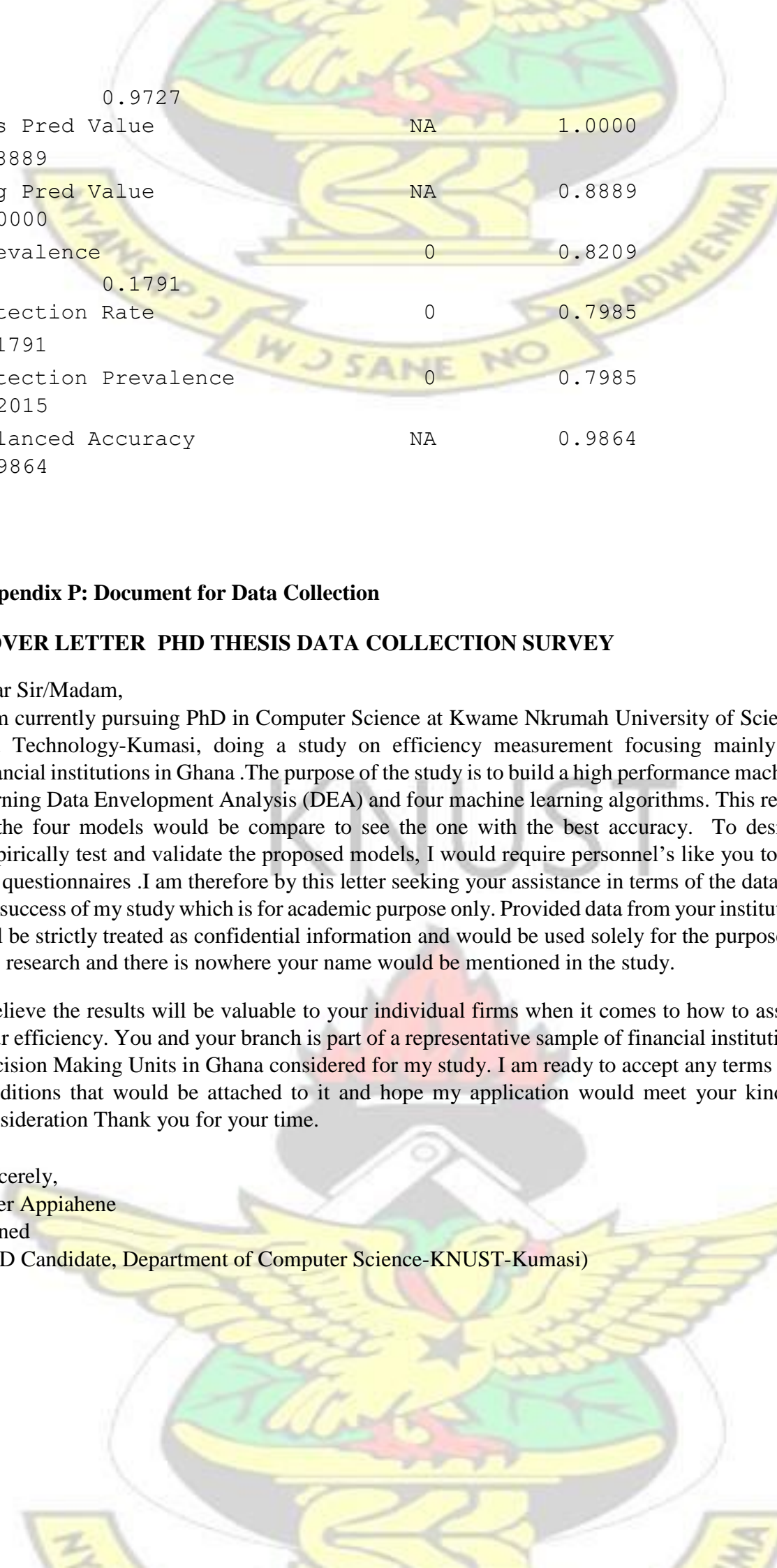
		Reference			
Prediction	Class 1	Class 2	Class 3	Class 4	
Class 1	0	0	0	0	
Class 2	0	107	0	0	
Class 3	0	0	0	0	
Class 4	0	3	0	24	

Overall Statistics

Accuracy : 0.9776
 95% CI : (0.936, 0.9954)
 No Information Rate : 0.8209
 P-Value [Acc > NIR] : 1.48e-08
 Kappa : 0.9274
 McNemar's Test P-Value : NA

Statistics by Class:

	Class: Class 1	Class: Class 2	Class:
Class 3	Class: Class 4		
Sensitivity	NA	0.9727	NA
1.0000			
Specificity	1	1.0000	



1	0.9727			
Pos Pred Value	NA	1.0000		NA
0.8889				
Neg Pred Value	NA	0.8889		NA
1.0000				
Prevalence	0	0.8209		
0	0.1791			
Detection Rate	0	0.7985		0
0.1791				
Detection Prevalence	0	0.7985		0
0.2015				
Balanced Accuracy	NA	0.9864		NA
0.9864				

Appendix P: Document for Data Collection

COVER LETTER PHD THESIS DATA COLLECTION SURVEY

Dear Sir/Madam,

I am currently pursuing PhD in Computer Science at Kwame Nkrumah University of Science and Technology-Kumasi, doing a study on efficiency measurement focusing mainly on financial institutions in Ghana .The purpose of the study is to build a high performance machine learning Data Envelopment Analysis (DEA) and four machine learning algorithms. This result of the four models would be compare to see the one with the best accuracy. To design, empirically test and validate the proposed models, I would require personnel's like you to fill out questionnaires .I am therefore by this letter seeking your assistance in terms of the data for the success of my study which is for academic purpose only. Provided data from your institution will be strictly treated as confidential information and would be used solely for the purpose of this research and there is nowhere your name would be mentioned in the study.

I believe the results will be valuable to your individual firms when it comes to how to assess your efficiency. You and your branch is part of a representative sample of financial institutions Decision Making Units in Ghana considered for my study. I am ready to accept any terms and conditions that would be attached to it and hope my application would meet your kindest consideration Thank you for your time.

Sincerely,

Peter Appiahene

Signed

(PhD Candidate, Department of Computer Science-KNUST-Kumasi)

